Unit Selection and Waveform Concatenation Strategies in Cantonese Text-to-speech

Oey Sai Lok

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Philosophy
in
Electronic Engineering

August 2005

Abstract of thesis entitled:

Unit Selection and Waveform Concatenation Strategies

in Cantonese Text-to-speech

Submitted by **Oey Sai Lok**

for the degree of **Master of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong**

in **August 2005.**

Cantonese is one of the Chinese dialects widely used in southern China. Like other Chinese dialects, Cantonese is a monosyllabic language. Cantonese Text-to-Speech (TTS) technology has been researched for some years. However, the performance of the synthesized speech is still not in high standard. Intelligibility and naturalness are the major concerns. Previous work on sub-syllable based Cantonese TTS showed that the naturalness can be improved by retaining the co-articulation between syllables. But the overall performance is not good. In this research, we focus on the acoustic synthesis technique for a corpus-based concatenative Cantonese TTS system. We address two major problems, namely the selection of the speech segments for concatenation and the methods used to concatenate them.

The concatenation of two speech segments is done in two steps: 1) determining the concatenation points for the selected speech segments; 2) concatenating the waveforms and smoothing. Based on the acoustic-phonetic properties, the Cantonese phonemes are classified carefully into different groups. Different approaches of concatenation are proposed for these groups.

We propose a unit selection scheme to fully utilize all available units in the acoustic inventory of sub-syllable units. The main idea is to allow the largest number of candidates and select the best one from the acoustical point of view. We define different levels of similarity between the candidate segments. At different levels, we may find multiple candidates for a desired unit. The most appropriate candidate is selected based on the match of acoustic properties with the prosody requirements.

Objective test is carried out to check the acoustic difference between the selected segments and the target prosody. It shows that segments selected by the proposed scheme are better matched with the target prosody than the baseline system. Subjective listening test is carried out to collect responses from human users. Results show that the proposed methods outperform the baseline system with a significant increase of mean opinion score (MOS).

摘要

粵語是中國南方最常用的方言。與其他中國方言一樣，它是一種單音節語言。粵語語音合成系統已發展了一段時間，但是合成語音的效果始終不夠理想，關鍵問題在於合成語音的清晰度和自然度。利用子音節作爲基本合成單元的系統已被證實在自然度方面有一定的改善，但總體的表現仍未能令人滿意。本論文主要討論的問題是基於拼接技術的粵語語音合成系統中，「聲學合成」部分的研究。主要的考慮點在於 1) 如何爲拼接選擇適當的音段，以及 2) 不同音段的拼接方法。

在音段拼接問題上，我們需要考慮每個音段的拼接點及其拼接的方法。此論文利用粵語聲學和語音學的特徵，將粵語的所有音素作出分類，對於不同的種類提出不同的拼接方法。

我們設計了一個音段選擇機制，可以充分利用所有儲存在語音庫內的音段。目的在於提供盡可能多的後備音段作爲選擇，並從聲學的角度選出最合適的一個作拼接之用。對於眾多的後備音段，我們首先定義了不同的相似程度。在每一個相似程度上，我們都可能爲一個目標音段找出幾個候選單元。基於韻律要求，我們從中選出一個與其聲學特徵最相似的音段。

客觀的測試方法是將被選擇的音段與其聲學模型作比較，結果顯示新的選擇音段機制能更有效地選出符合目標單元聲學特徵的音段作拼接之用。主觀測聽試驗中，我們收集聽者對系統的評分，並作出分析。結果顯示新的聲學合成方法能夠提高其合成語音的清晰度及自然度。

Acknowledgement

I am greatly indebted to Prof. Tan Lee for his supervision and support throughout this research. I also wish to give my thanks to Prof. P. C. Ching, Prof. Y. T. Chan and Prof. William S-Y. Wang for their valuable suggestions. I would like to thank Ms. Y. J. Li and Ms. Y. Qian for their knowledge sharing and Mr. Arthur Luk for his technical support.

I would like to thank all the colleagues and friends in DSP-ST laboratory for giving me enjoyable and inspiring working environment. I would also thank all the participants for their feedback and comments on my work. To name only some of them: Mr. Dexter Chan, Ms. Joyce Chan, Ms. Yvonne Lee, Dr. James W. Minett, Mr. Raymond Ng, Ms. L. Y. Ngan, Mr. H. Ouyang, Mr. C. Qin, Ms. W. Y. Tsui, Mr. M. Yuan, Mr. N. H. Zheng and Ms. Y. Zhu. Further, I would like to thank the undergraduate students in helping me to run the perceptual test. They include Mr. C. K. Chan, Mr. C. K. Cheung, Ms. Joyce Ho, Mr. K. K. Lam, Mr. C. T. Li, Mr. K. M. Poon, Mr. Arthur Yu and Mr. Wilson Yu.

Finally, I would like to express sincere gratitude to my parents for their patience, support and understanding throughout this research.

Contents

List of Tables

List of Figures

Chapter 1
Introduction

In our daily life, speech and text are the most common ways of communication and information exchange, both between human beings and between humans and computers. Compared with writing, speaking is more convenient and efficient because it does not require additional apparatus. On the other hand, large amount of information are stored as text.

Computers are very popular nowadays. Human-computer interaction becomes more and more important in this information age. There are a lot of applications that require frequent and natural interaction between humans and computers. An automatic conversion system from textual information to human speech is one of the key elements for human-computer communication, especially for the visually impaired who cannot read printed text.

There are two issues of concern in developing a Text-to-Speech (TTS) system, namely intelligibility and naturalness. Intelligibility means that human listeners can easily acquire the content of the synthesized speech. Existing TTS systems do not have many problems in achieving a high level of intelligibility. On the other hand, naturalness refers to the degree that the synthesized speech sounds like natural human speech. To achieve high naturalness is still a challenge for TTS research. For most commercial systems, it is not difficult to tell that their outputs are produced by machines instead of spoken by humans. In recent years, scientists have put extensive efforts on making synthesized speech sound as natural as possible.

1.1 An overview of Text-to-Speech technology

TTS refers to the automatic conversion process from input text to acoustic speech signals. The input text is composed of symbols commonly used by the human society to exchange information. It is language dependent. For example, alphabets are used in Western languages such as English while characters are used in Chinese. These alphabets and characters may not have direct pronunciations. Acoustic speech signals generated by a TTS system are sound signals which can be understood by human listeners. When people are using TTS system, it is expected that the system could interact like communicating with others using text and voice directly.

Basically, a TTS system is composed of three parts, namely text processing, acoustic synthesis and prosody control. Figure 1.1 shows the system flow. Each module will be discussed in detail.

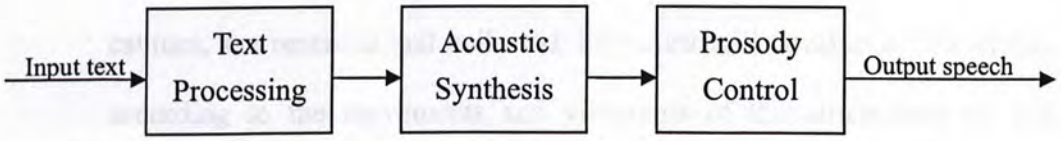


Figure 1.1 Overview of a Text-to-Speech system

1.1.1 Text processing

There are various methods of inputting text into a computer, such as optical scanning, keyboard typing or simply from a pre-stored file. The primary function of text processing is to convert the input text into a sequence of pronunciation symbols that can be spoken (or "synthesized"). Each of these symbols typically corresponds to a

basic sound unit. For example, the pronunciation of Chinese character "啊" is represented by the symbol /aa/. Text with special format and meaning, such as numbers, date and time format, etc, are also acceptable as the input in text processing stage.

1.1.2 Acoustic synthesis

The sequence of pronunciation symbols generated by text processing will be passed to the acoustic synthesis module to generate synthetic speech. Several methods are used to synthesize the speech. These include articulatory approach, parametric approach and corpus-based approach.

1) Articulatory approach

This approach attempts to generate speech signals by replicating the process of human speech production. Movement and vibration of different articulators, including vocal cords, palate, tongue and lip, oral and nasal cavities, are recorded and collected. Scientists will develop a TTS system according to the movements and vibrations of the articulators [1] [2]. However, it is quite difficult to collect such information. Only a few synthesis systems have been developed using this approach and their performance is relatively poor as compared to those using other approaches [3].

2) Parametric approach

In parametric synthesis system, speech signal is synthesized based on rules and pre-trained parameters. The parameters may include pitch level,

duration, intensity, formant positions and bandwidths, etc. This approach generally provides acoustically accurate speech output and meets the basic requirements of correct pronunciation and prosody. However, the synthesized speech usually sounds unnatural. People always comment it as "machine speech".

3) Corpus-based approach

In this approach, speech output is produced from a huge amount of speech data that are recorded and processed carefully beforehand. To synthesize a spoken sentence, the most appropriate speech segments from the database are found and concatenated. Theoretically, speech segments could be of any size, including sentence, word, sub-word and phoneme. However, sentence and word-level units are impractical simply because there are too many of them. For example, there are more than 1700 different words for a single Chinese character "不" in the lexicon we are using in Cantonese TTS. As a result, segments at sub-word level are more preferable for corpus-based TTS. To better capture the co-articulation effect, the important contextual variations of the sub-word units are often included. Sometimes segments of variable-length can be used. For example, the corpus may include some commonly used words, in addition to the sub-word units. In general, corpus-based systems can produce speech of fairly good quality.

1.1.3 Prosody modification

Generally, prosody refers to the properties of a continuous speech such as pitch and fundamental frequency (F0), loudness, rhythm and tempo [4]. It is said to be

supra-segmental phenomenon because such events can be aligned with phonemes, syllables or groups of syllables [5]. It serves to structure the flow of the speech including the modifications in stress and intonation [6].

The speech produced by the acoustic synthesis module may not carry the desired prosody. For the corpus-based approach, the coverage of the acoustic inventory is always limited. Even the best available units may not match the target F0 profile, duration, etc. Then there is a need for prosody modification in which the prosody-related acoustic properties are changed with respect to certain specified targets. Given the linguistic content of an utterance, the prosody targets are predicted by a prosody model. The prediction can be either based on a set of rules or the results from statistical analysis of natural speech [7] [8].

1.2 Trends in Text-to-Speech technologies

Many researchers have put a lot of efforts on developing TTS systems synthesizing speech with high standard of intelligibility and naturalness, such as the Festival TTS by the University of Edinburgh [7], IBM Cantonese TTS system [8] and Microsoft Research MSRA Text to Speech [9]. For example, the IBM Cantonese TTS system uses a large-corpus that contains about 2200 sentences with the length varying from 10 to 30 Chinese syllables. The corpus-based approach is commonly used [10] [11]. It can provide a relatively high standard of synthesized speech, in terms of intelligibility and naturalness, when compared with other approaches. Our research on Cantonese TTS also adopts this approach.

When we are developing a TTS system using corpus-based approaches, two issues

need to be considered. Firstly, the size of the acoustic inventory and the speech data required for the corpus are determined. Using a huge inventory can allow you to find the most suitable speech segments among all the available speech segments in the inventory. However, it requires a lot of manpower in developing such a huge database, especially for collecting and recording speech data, and for processing the recorded speeches. Small inventory requires a relatively smaller development cost in processing job and storage, but the availability of the selection of segments may be reduced due to the small number of duplicated copies of the segments.

Another issue in deciding a corpus-based TTS system is the segmental level used in speech segments. High-level segments, such as sentence based or word based, can provide us a better match in speech context as a longer speech segments are from the same speech context in a recorded wave. Also, it requires less concatenation points within a speech utterance. However, it is not possible for us to guarantee that the speech corpus contains all sentences and words in any patterns. So low-level segments, such as sub-word based or syllable based, are preferable. It is more practical if the size of the inventory is relatively small. However, smaller in speech segments may come from different recording contexts. Also, it will introduce more concatenation points between segments and may have a higher chance in generating artifacts.

Due to the above considerations, our TTS system will use a relatively large corpus size but the selection of the speech segments are done in syllable levels. Such an approach can allow us to have choices of the speech segments within the inventory. We can make use of the properties which will be discussed in the latter chapters to select the most appropriate segments for synthesizing speech. Since this approach requires a larger amount of concatenation points, so another critical point in our study

is to propose a method in determining the concatenation points in order to avoid artifacts.

1.3 Objectives of this thesis

The main objective of this thesis is to apply acoustic and phonetic properties in Cantonese to improve the intelligibility and naturalness of the synthesized speech. We targeted on the synthesis stage in TTS system only and assume other stages of the system are working in good conditions and no modification is applied.

There are two major parts in acoustic synthesis stage. They are 1) selecting the most suitable speech segments for synthesis and 2) using the proposed methods to concatenate selected speech segments together. We will improve the system by the following ways:

1) To develop concatenation strategies that can exploit the different acoustic and phonetic properties of different speech units;

2) To develop a set of criteria in selecting the speech segment (among many choices) that is the most appropriate for concatenation.

After implementing the suggested strategies and methods mentioned in this thesis, we believe the speech synthesizer can select suitable speech segments for synthesizing speeches. The acoustic and phonetic properties of the selected speech segments are close to the target prosody so that the overall perceptual quality is improved in both the intelligibility and naturalness.

This thesis will mainly focus on the Cantonese synthesizer. But we believe that the methods proposed in the latter chapters are applicable to other synthesizers using different approaches in different languages. There are several reasons supporting us as listed below:

1) We believe that the concatenation strategies proposed are applicable to all systems using concatenative and corpus-based approach. Our concatenation strategies mainly base on the acoustic and phonetic properties of the speech signals. These properties are not restricted to Cantonese or other Chinese languages only. Other languages may share similar behaviour like Cantonese. Our proposed concatenation methods may help them in concatenating different types of units together.

2) Unit selection process is applicable in all TTS systems when more than one available speech segment is found inside the corpus. Selection process does not have a direct relationship to the size of the speech corpus you used. Once you have choices in different speech segments, you will undergo a selection process to choose the most appropriate segment. The selection process will mainly focus on the acoustic properties of the speech segments including pitch levels, duration length and intensity levels. These parameters are available in all spoken languages. So the proposed method is language-independent.

1.4 Outline of the thesis

In the next chapter, we will focus on the spoken language we synthesize – Cantonese. Basic information and phonology of Cantonese will be introduced. Finally, a brief introduction in acoustic-phonetic properties of Cantonese syllables will be mentioned.

Chapter 3 will focus on the background of Cantonese TTS system. First of all, we will have a general overview of a Cantonese TTS system. After that, we will have a description on the two systems using different approaches in synthesizing speech. We will pay attention to the acoustic inventory and the speech synthesis process of the two systems. Finally, we will focus on the problems found inside the systems currently being used.

Chapter 4 will focus on the concatenation strategies proposed for the sub-syllable based concatenative Cantonese TTS system. Previous works in waveform concatenation are introduced. After that, problems and difficulties in concatenation are described. Grouping of different units will be discussed and detailed concatenation methods are proposed including the determination of the concatenation points and the method used in concatenating two speech segments. Selected examples in concatenating different types of units will be shown at the end of this chapter.

Chapter 5 will focus on the newly introduced unit selection process in our TTS system. Assumptions and feasibility discussion are made in the first part of this chapter. Following that we will address the factors that are being used in unit selection process. Such factors may apply to different situation in Initial and Final units. Two parts in unit selection process are discussed including the selection of the candidate units and

the factors used in selecting the best segment among candidate units. Different process will be held according to the availability of the units and the groupings of the units. Selected examples in selecting different types of units are shown at the end of this chapter.

Chapter 6 will present the results in performance evaluation. Two experiments are done based on a list of test words and paragraphs including an objective test on acoustic parameters and a subjective perceptual test. In objective test, different acoustic parameters including pitch and duration are measured and compared with the target values provided by the system. For the subjective perceptual test, we will invite subjects who are Cantonese native speakers to rate the system according to the synthesized speeches they heard. Results of the two testes are provided with a brief analysis at the end of each test. A summary is concluded at the end of this chapter.

Lastly, conclusions are to be drawn in chapter 7 and several suggested future works are included in this chapter for reference.

References:

[1] M. Mohan Sondhi & Daniel J. Sinder, "Articulatory modeling: a role in concatenative text to speech synthesis", in Text to Speech Synthesis: New Paradigms and Advances, Shrikanth Narayanan & Abeer Alwan, Prentice Hall Professional Technical Reference, 2005, pp. 63 – 88.

[2] Manfred R. Schroeder, "Speech Synthesis", in Computer Speech: Recognition Compression Synthesis, Manfred R. Schroeder, Springer, 2004, pp. 129 – 134.

[3] B. Gabioud, "Articulatory models in speech synthesis", in Fundamental of speech synthesis and speech recognition, E. Keller, John Wiley & Sons, pp. 215-230.

[4] Y. J. Li, "Prosody Analysis and Modeling for Cantonese Text-to-Speech", M. Phil. Thesis, the Chinese University of Hong Kong, 2003

[5] T. Dutoit, "Automatic Prosody Generation", in an Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, 1997, pp. 129 - 176

[6] S. Werner & E. Keller, "Prosodic Aspects of Speech", in Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges, Eric Keller, John Wiley & Sons, 1994, pp. 23 – 40.

[7] J. van Santen, J. Wouters & A. Kain, "Modification of speech: a tribute to

Mike Macon", in Proc. Speech Synthesis, Santa Monica, CA, USA, 2002, pp. 1 – 6.

[8] Tan Lee, Helen Meng, W. Lau, W. K. Lo & P. C. Ching, "Micro-prosodic control in Cantonese Text-to-Speech synthesis", in Proc. Eurospeech-99, Vol. 4, pp. 1855-1858, Budapest.

[9] http://www.cstr.ed.ac.uk/projects/festival/, homepage of the Festival TTS.

[10] H. Li, F. Chen, L.Q. Shen & X, J, Ma, "Trainable Cantonese/English Dual Language Speech Synthesis System", in Proc. ICASSP vol. 1, Hong Kong, 2003, pp 508-511

[11] https://www.research.microsoft.com/speech/tts.asp, homepage of the Microsoft Research MSRA Text to Speech.

[12] M. Kitai, K. Hakoda & S. Sagayama, "Trends of ASR and TTS applications in Japan", in Proc. Interactive Voice Technology for Telecommunications Applications, Basking Ridge, NJ, USA, 1996, pp 21 - 24.

[13] J. Schroeter, A. Conkie, A, Syrdal, M. Beutnagel, M. Jilka, V. Storm, Yeon-Jun Kim, Hong-Goo Kang & D. Kapilow, "A perspective on the next challenges for TTS research", in Proc. Speech Synthesis, Santa Monica, CA, USA, 2002, pp. 211 – 214.

Chapter 2
Cantonese Speech

2.1 The Cantonese dialect

Cantonese is one of the major dialects in China. It is commonly used in the Southern part of China, especially in the provinces of Guangdong (廣東) and Guangxi (廣西), as well as Hong Kong and Macau. At the same time, Cantonese is frequently used in many overseas Chinese communities, e.g. Singapore, Malaysia, Canada and United States. Statistics show that more than 60 million of the world's total populations speak Cantonese [1]. It ranks among the top 20 most popular spoken languages in the world [2] [3]. In mainland China, it is the third most popular dialect used [2] just after Mandarin (官話漢語 / 普通話) and Wu (吳語) dialects.

There are several variations in Cantonese dialects including Yuehai (粵海) (sometimes called Guangfu (廣府) or Hong Kong Cantonese), Siyi (四邑), Gaoyang (高陽) and Guinan (桂南) [2]. Their pronunciations are similar among different variations. In this thesis, our discussion will base on the most commonly used Cantonese, which is the Hong Kong Cantonese. All the pronunciations used in this thesis are based on the Hong Kong Cantonese.

Cantonese is a monosyllabic and tonal language. It has a relatively simple phonemic structure. Each Chinese character is pronounced as a syllable. Each syllable unit carries a specific lexical tone. Cantonese is homophonic. Homophone means that different Chinese characters can have the same pronunciations. An example is the

syllable /fu3/. It is the common pronunciation of the Chinese characters "褲" (pants), "富" (rich), "副" (supplement) and "庫" (treasury). On the same time, Chinese is polyphonic characters. It means that a Chinese character may have multiple pronunciations. For example, the character "樂" can be pronounced as /lok6/ (happiness), /ngok6/ (music), /lok3/ (a surname) and /ngaau6/. In short, the mapping between Chinese characters and Cantonese pronunciations are not one-to-one.

2.2 Phonology of Cantonese

As said above, each Cantonese syllable carries a lexical tone. If the tone is not considered, the syllable is referred to as a base syllable. Each base syllable can be further divided into two parts: Initial and Final. The Initial consists of an onset consonant and the Final consists of a vowel nucleus and a consonant coda. The basic structure of a Cantonese syllable is illustrated as in Figure 2.1.

<table>
<tr><td colspan="4">Cantonese syllable</td></tr>
<tr><td colspan="3">Base syllable</td><td rowspan="3">Tone information
(required)</td></tr>
<tr><td>Initial unit</td><td colspan="2">Final unit</td></tr>
<tr><td>Consonant onset
(optional)</td><td>Vowel nucleus
(required)</td><td>Consonant coda
(optional)</td></tr>
</table>

Figure 2.1 Basic structure of the Cantonese syllable.

Different Romanization systems have been used to label Cantonese phonemes and syllables. They include Yale [4], Sidney Lau [5] and IPA system [6]. In this thesis, we use the Jyutping system, which was developed by the Linguistic Society of Hong Kong [7].

2.2.1 Initials

There are a total of 20 Initials in Cantonese. They include the null Initial, which means that there is no Initial. According to their manner of articulation, the Initials are categorized as shown in Table 2.1.

Types of initial	Manner of articulation
Plosive and affricate	Voice produced by stopping the airflow using the lips and teeth, and followed by a sudden release of air.
Fricative	Voice produced by the friction in the narrow opening of the vocal tube.
Voiced unit (including nasal, semi-vowel and liquid)	Voice produced by the resonance of the vocal cords.

Table 2.1 Phonetic properties of different Initial unit types.

Table 2.2 gives an example character and a syllable for each Initial of Cantonese.

Vocal cord vibration	Manner of articulation	LSHK symbol	Example		
			Transcription	Meaning in Chinese	Meaning in English
Unvoiced	(null)	-	/aa1/	啊	(exclamatory particle)
	Plosive	b	/baa1/	爸	father
		d	/daa2/	打	beat
		g	/gaa1/	家	home
		p	/paa3/	怕	fear
		t	/taa1/	他	he
		k	/kaa1/	卡	card
		gw	/gwaa3/	掛	hang
		kw	/kwaa1/	誇	exaggerate
	Affricate	c	/caa1/	差	differ
		z	/zaa3/	炸	fry (in oil)
	Fricative	f	/faa1/	花	flower
		s	/saa1/	沙	sand
		h	/haa1/	哈	sound of laughter
Voiced	Nasal	m	/maa1/	媽	mother
		n	/naa4/	拿	take
		ng	/ngaa4/	牙	tooth
	Glide / Semi-vowel	j	/jaa6/	廿	twenty
		w	/waa1/	娃	doll
	Liquid	l	/laa3/	喇	horn

Table 2.2 Examples in different Initials.

2.2.2 Finals

There are 53 Finals in Cantonese. Their composition can be as simple as having a single long vowel and as complicated as containing a vowel nuclei followed by a nasal or stop consonant. Table 2.3 lists all of the 53 Cantonese Finals.

Vowel nucleus	Null coda	Vowel coda		Nasal coda			Stop coda		
	-	i	u	m	n	ng	p	t	k
-	-	-	-	m	-	ng	-	-	-
a	-	ai	au	am	an	ang	ap	at	ak
aa	aa	aai	aau	aam	aan	aang	aap	aat	aak
e	e	ei	-	-	-	eng	-	-	ek
eo	-	eoi	-	-	eon	-	-	eot	-
i	i	-	iu	im	in	ing	ip	it	ik
o	o	oi	ou	-	on	ong	-	ot	ok
oe	oe	-	-	-	-	oeng	-	-	oek
u	u	ui	-	-	un	ung	-	ut	uk
yu	yu	-	-	-	yun	-	-	yut	-

Table 2.3 Finals combination table consists of vowel nucleus and codas.

Table 2.4 gives an example character and a syllable for each Final of Cantonese.

LSHK symbol	Example			LSHK symbol	Example		
	Transcription	Chinese meaning	English meaning		Transcription	Meaning in Chinese	Meaning in English
aa	/saa1/	沙	sand	in	/cin1/	千	thousand
aai	/faai3/	快	fast	ing	/ming4/	明	bright
aau	/baau1/	包	pack	ip	/zip3/	接	connect
aam	/saam1/	三	three	it	/sit6/	舌	tongue
aan	/saan1/	山	hill	ik	/zik6/	直	straight
aang	/laang5/	冷	cold	o	/po3/	破	break
aap	/daap3/	答	answer	oi	/oi1/	哀	sad
aat	/baat3/	八	eight	ou	/sou3/	數	number
aak	/baak6/	白	white	on	/on1/	安	peaceful
ai	/sai1/	西	west	ong	/gong1/	江	river
au	/sau1/	收	collect	ot	/got3/	割	cut
am	/sam1/	心	heart	ok	/dok6/	度	measure
an	/gan1/	根	root	oe	/hoe1/	靴	boots
ang	/wang4/	宏	wide	oeng	/soeng2/	想	think
ap	/sap1/	濕	wet	oek	/goek3/	腳	foot
at	/bat1/	不	not	u	/fu1/	夫	husband
ak	/bak1/	北	north	ui	/bui1/	杯	cup
e	/se4/	蛇	snake	un	/wun6/	換	change
ei	/sei3/	四	four	ung	/zung1/	中	central
eng	/beng2/	餅	biscuits	ut	/kut3/	括	include
ek	/sek6/	石	stone	uk	/fuk1/	福	fortune
eoi	/deoi3/	對	correct	yu	/syu1/	書	book
eon	/leon6/	論	discuss	yun	/zyun3/	轉	move
eot	/leot6/	率	rate	yut	/syut3/	說	speak
i	/si5/	市	market	m	/m4/	唔	(colloquial expression)
iu	/siu2/	小	small				
im	/dim2/	點	dot	ng	/ng5/	五	five

Table 2.4 Examples in different Finals.

2.2.3 Tones

Cantonese is rich in tones. There are altogether nine different tones in Cantonese. For the convenience of computer processing, they are labeled by numerals from 1 to 9, as shown in Figure 2.2. The first six tone classes are called non-entering tones while the last three are entering tones.

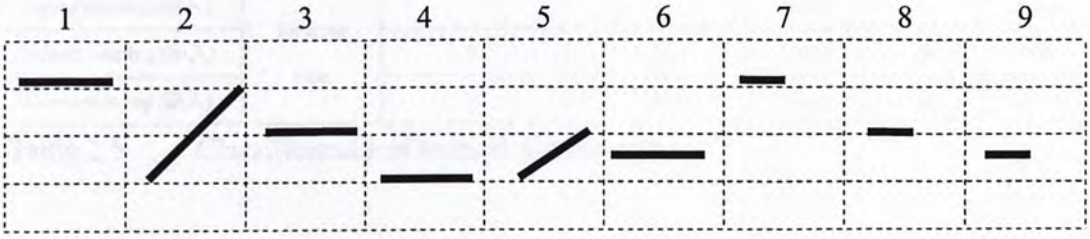


Figure 2.2 Nine tone classes in Cantonese.
(Horizontal length: Duration of the syllable;
Vertical level: Pitch levels in frequency.)

Syllables with entering tones have two special properties. First of all, their duration is relatively short as compared with syllables with the non-entering tones. Secondly, their vowel nucleuses are always followed by a stop coda /p/, /t/ or /k/. Actually, the pitch profiles of the entering tones are very similar to those non-entering tones. We can combine these three entering tones to other six non-entering tones. We can distinguish them easily by the presence of the stop coda. Groupings of tones are shown in table 2.5. In this thesis, we will use the six-tone representation.

Tone representation	Groupings / Classifications			Example		
	Entering / non-entering tone?	Tone class number (nine-tone system)	Tone class number (six-tone system)	Syllable	Chinese meaning	English Meaning
Upper level (陰平)	Non-entering tone	1	1	fu1	夫	husband
Upper rising (陰上)		2	2	fu2	苦	bitter
Upper departing (陰去)		3	3	fu3	富	rich
Lower level (陽平)		4	4	fu4	扶	hold on
Lower rising (陽上)		5	5	fu5	婦	wife
Lower departing (陽去)		6	6	fu6	付	give
Upper entering (陰入)	Entering tone	7	1	fuk1	福	fortune
Middle entering (中入)		8	3	fut3	闊	wide
Lower entering (陽入)		9	6	fuk6	伏	crouch

Table 2.5 Classifications of tones in Cantonese.

2.3 Acoustic-phonetic properties of Cantonese syllables

Phonetics is a study of speech sounds. The study can be from articulatory, auditory and acoustic points of view. Articulatory phonetics concerns how speech sounds are produced by the human articulatory organs. Different phonetic units are classified according to their articulation gestures. An example has been given as in Section 2.2.1 (Initials). On the other hand, auditory phonetics is a study of how speech sounds are perceived by human hearing organs. It focuses on the mechanism in how human processed a speech signals.

Acoustic-phonetics is a scientific study of sound wave made by human vocal organs. It uses measurable parameters to represent a sound unit. Acoustic-phonetic studies provide us with accurate description of a spoken utterance by its acoustic properties, including intensity, duration, pitch, formants. In this section, we will study the acoustic-phonetic properties of Cantonese syllables.

Figure 2.3 shows the waveform and wideband spectrogram of a sequence of Cantonese syllables. The boundaries of Initial and Final units are also given.

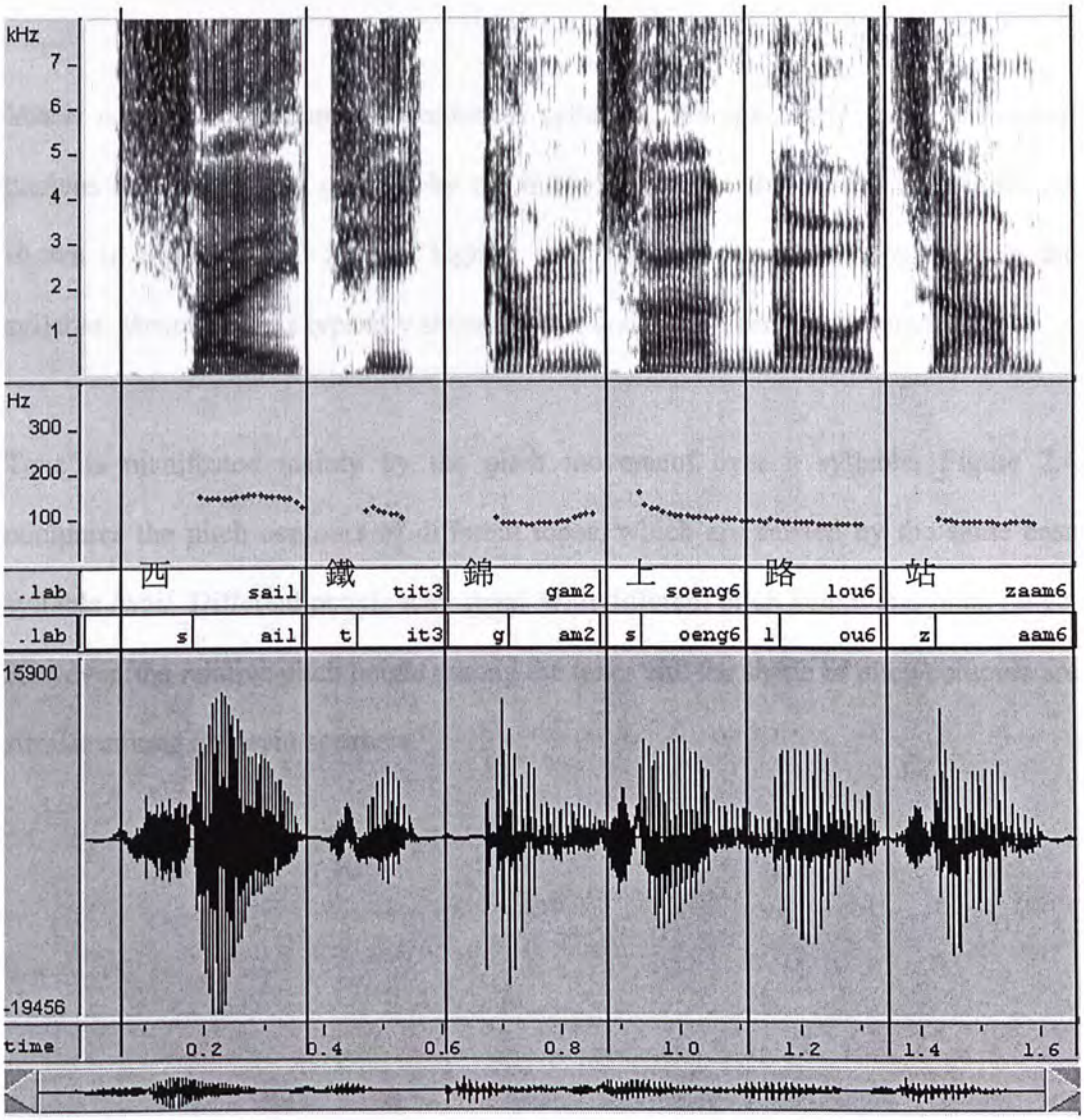


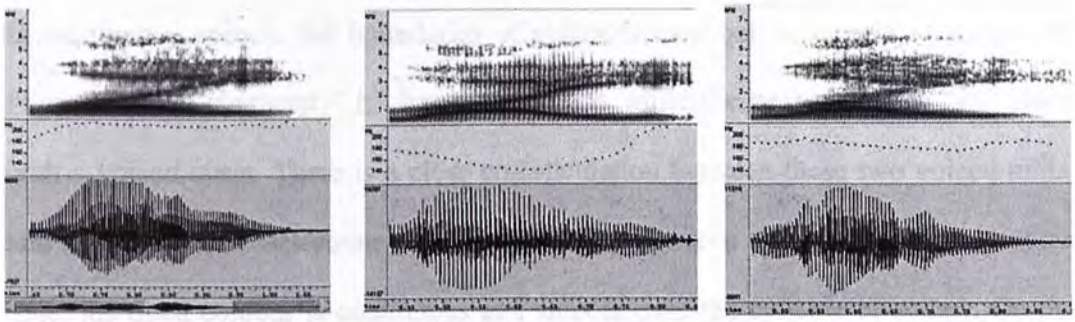
Figure 2.3 A Cantonese speech utterance with three displays (waveform, spectrogram and pitch contours).

Cantonese syllables have a general structure of consonant-vowel-consonant, where the consonant onset and the consonant coda are optional. The onset consonants may be voiced or unvoiced. The formant structure of a voiced consonant onset may not be

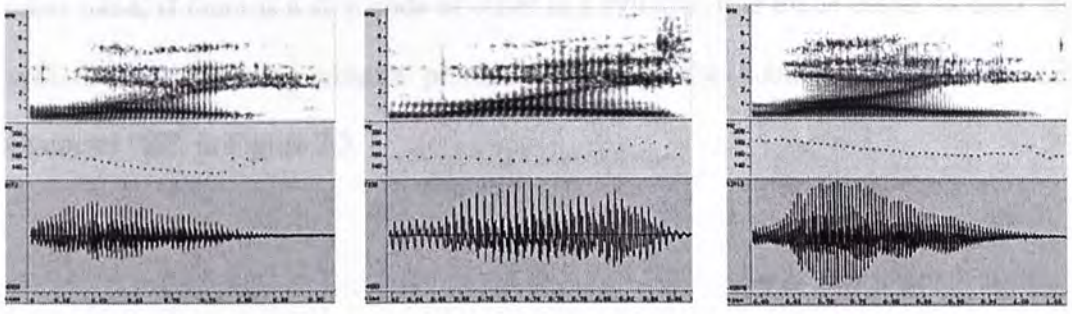
as clear as that of a vowel. As can be seen in Figure 2.3, the consonant onsets are transient segments. Their durations are relatively short when compared with the vowel nuclei. Their intensity levels are much weaker as compared with the other parts in the syllable.

Vowel nuclei are the cores of Cantonese syllables. We can easily locate the vowel nucleus of a Cantonese syllable by the intensity peak in the speech waveform. As shown in Figure 2.3, it has the highest intensity level among all segments in the syllable. Vowel nucleus typically shows a clear and stable formant structure.

Tone is manifested mainly by the pitch movement over a syllable. Figure 2.4 compares the pitch contours of different tones, which are carried by the same base syllable /wai/. Different people may speak with different pitch height and pitch range. However, the relative pitch height among the tones and the shape of pitch contours are similar among different speakers.



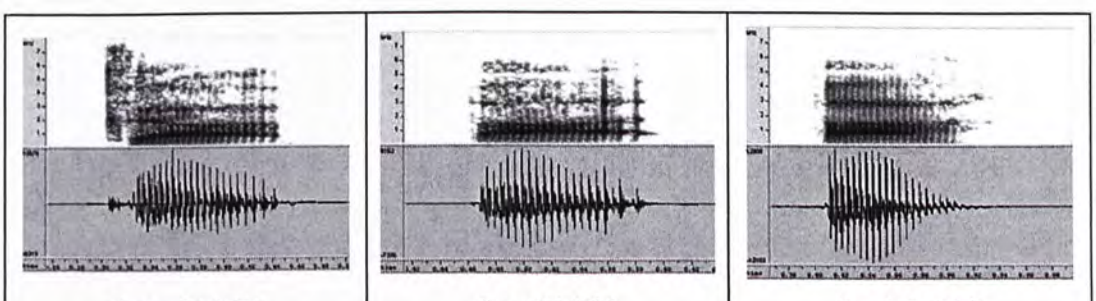
/wai1/ (威) Pitch: 209 Hz /wai2/ (委) Pitch: 143 Hz /wai3/ (喂) Pitch: 179 Hz



/wai4/ (維) Pitch: 139 Hz /wai5/ (偉) Pitch: 150 Hz /wai6/ (胃) Pitch: 164 Hz

Figure 2.4 Different pitch contours with the same base syllables.

There are two types of consonant codas in Cantonese, namely nasal and stop. Sometimes the ending part of the diphthongs like /ui/ and /aau/ are also considered to be a special type of coda. Stop codas in Cantonese have some special and interesting properties. These transient phonemes are not released. Its distinctive features are carried mainly by the preceding vowels. It is difficult to identify them by waveforms and spectrograms.



/gaap3/ (甲) /baat3/ (八) /baak3/ (百)

Figure 2.5 Spectrograms of three stop codas in Cantonese.

In continuous speech, the boundaries of syllables may not be clearly identified. In Figure 2.3, the character "上" has a nasal coda while the next character "路" starts with a voiced onset. There is a close co-articulation between these two voiced units, and it is difficult to determine an exact boundary between the syllables. At the same time, the pitch contour is continuous and smooth over the syllable boundary. On the other hand, if there is a stop coda or onset in a syllable, it is much easier to mark the syllable boundaries. A closure period is found at the boundary region, like the character "鐵" in Figure 2.3.

References:

[1] Ethnologue: Languages of the World, http://www.ethnologue.com/, Internet version, SIL International, 2005

[2] 維基百科，“粵語”, http://zh.wikipedia.org/w/index.php?title=%E7%B2%A4%E8%AF%AD&variant=zh, Internet version, 2005

[3] David P. Brown, Top 100 Languages by Population, http://www.davidpbrown.co.uk/help/top-100-languages-by-population.html, Internet version, 2004

[4] P. Huang, “Cantonese Dictionary”, New Haven: Yale University Press, 1970.

[5] S. Lau, “Elementary Cantonese”, the Government Printed, Hong Kong

[6] Sidney Lau, “Elementary Cantonese”, the Government Press, Hong Kong, 1972.

[7] Linguistic Society of Hong Kong (LSHK, 香港語言學學會), “Cantonese Romanization (粵語拼音方案)”, http://cpct92.cityu.edu.hk/lshk/Jyutping/Jyutping.htm, Internet version, 2002.

Chapter 3

Cantonese Text-to-Speech

3.1 General overview

For a Cantonese TTS system, the input text is a sequence of Chinese characters and the output is Cantonese speech. As described in Chapter 1, a typical TTS system can be divided into three modules, namely text processing, acoustic synthesis and prosody control. These modules in a Cantonese TTS system are explained below.

3.1.1 Text processing

The primary goal of text processing for Cantonese TTS is to convert the input sequence of Chinese characters into a string of pronunciation symbols. Given that a character can have a number of different pronunciations, the process of text processing is not as trivial as simple pronunciation assignment. The exact pronunciation of a character depends on its linguistic contexts. There may also be some special content in the input text that needs to be handled, for examples, abbreviations, alpha-numeric, date and time, punctuations. The output string of pronunciation symbols will be used for the synthesis of the desired sounds.

More precisely, text processing for Cantonese TTS involves two major steps:

1) Word segmentation:

   Word segmentation is a process to split a piece of input text into segments.

Each segment is a lexical word that can be found in the dictionary. Unlike English, written Chinese does not contain explicit delimiters that separate words. Word segmentation may be ambiguous. That is, there may be more than one way to segment a sentence.

2) Pronunciation assignment:

Segmented text will be assigned the desired pronunciations according to the segmented text given from the text segmentation. This is an important issue for languages that have polyphone characters. For example, the Chinese character "曾" has two different pronunciations - /cang4/ (a sign of past) and /zang1/ (a Chinese surname).

3.1.2 Corpus based acoustic synthesis

With the corpus based approach, Cantonese speech is generated by concatenating short speech segments. Given the huge number of different Chinese words, it is impractical to use word as the basic concatenation unit. In our research, we focus on the concatenation of syllables and sub-syllable units. The acoustic inventory contains all these units and some of their contextual variations.

Typically, the text processing module generates a sequence of tonal syllables. The required speech segments that contain these syllables are then retrieved from the acoustic inventory. If the acoustic inventory is designed to contain units that are smaller than syllables, the syllable-level transcription needs to be further converted into Initial and Final units.

By concatenating the retrieved segments, a complete utterance is produced. The speech segments to be concatenated may come from different linguistic and acoustic contexts. Appropriate concatenation strategy is needed to make the segment transition as smooth as possible.

3.1.3 Prosodic control

Prosody refers to the properties of a continuous speech signals. Our prosodic controls focus on the modification of acoustic parameters of the synthesized speech.

A natural sentence carries some designated prosodic pattern, e.g. pitch movement, duration and intensity variation. Prosodic control in a TTS system attempts to approximate the natural prosody as much as possible. However, concatenated speech segments generally do not carry the desired prosody. This is due to that the speech segments are generally from many different utterances with different phonetic and prosodic contexts. Modification of prosodic parameters is required so that the output speech can carry the desired prosody.

There are two issues to be considered in prosodic control. Firstly, we need to specify the target prosody. Secondly, we need an effective signal processing technique by which the prosody-related acoustic parameters can be modified without much deteriorates the speech quality.

1) Prosody prediction:

From the input text we can predict different prosodic parameters including fundamental frequency (F0), duration and intensity. Each language has its

own prosodic properties, which can be observed by analyzing natural human speech. Usually the prediction is based on a set of rules or the results from statistical analysis of natural speech. The Cantonese TTS systems developed by our laboratory use the results of statistical analysis from a large speech Cantonese database [1]. The analysis was focused mainly on the syllable-wide pitch contour, syllable's duration and pause length between successive syllables.

2) Prosodic modification:

We can modify the prosody of the synthesized speech utterances according to the prosodic parameters generated in prosody prediction. Time-domain pitch-synchronous-overlap-add (TD-PSOLA) is used for prosodic modification because of its simplicity in algorithm and good quality of output speech [2] [3]. In TD-PSOLA, original speech signal is transformed into a stream of overlapping short-time signal which is obtained by a sequence of pitch-synchronous windows. This stream of original short-time signal will be modified to produce a sequence of synthesis short-time signal. The modification is done according to the prosodic parameters provided by the results from prosody prediction. Finally, the synthesized speech is obtained by overlap-adding these synthesis short-time signals together.

3.2 Syllable based Cantonese Text-to-Speech system

Given the monosyllabic nature of Cantonese, it is straightforward to use syllables as the basic synthesis unit for Cantonese TTS. Such an approach has also been used in many Mandarin TTS systems [4]. In a syllable based system, all tonal syllable units

need to be included in the acoustic inventory. The output speech is produced by concatenating the retrieved segments at the syllable boundaries.

A Cantonese TTS system was developed in our laboratory using syllable-based synthesis units [5]. The acoustic inventory contains 1,800 tonal syllables, which cover most of the pronunciations in Cantonese. Although these speech segments are accurate in terms of phonetic compositions and tones, the linguistic and acoustic contexts from which they are extracted are highly variable. Prosodic modification is applied with the TD-PSOLA technique. The target prosody is specified according to the results of statistical analysis as described in Section 3.1.3.

3.3 Sub-syllable based Cantonese Text-to-Speech system

The syllable based system could produce Cantonese speech with fairly good intelligibility. However, the naturalness of speech was found to be quite poor [6]. The transitions between consecutive syllables are hard and abrupt. Perceptually, each syllable can be recognized clearly but the syllables are all isolated. To better capture the inter-syllable transition, a sub-syllable based Cantonese TTS system was developed in our laboratory [6].

3.3.1 Definition of sub-syllable units

A sub-syllable unit is defined to contain two parts. Each part is either an Initial or a Final. Therefore, there are two different types of sub-syllable units: Initial-Final (I-F) and Final-Initial (F-I) units. An I-F unit is essentially a monosyllable. An F-I unit is an inter-syllable unit, with the Final and the Initial coming from two different syllables.

Since the Initial is optional in some of the syllables, the inter-syllable units may also be in the form of Final-Final, with the two Finals from different syllables. In addition, Initials and Finals used at the beginning and the ending of a sentence are considered specially (stand-alone units). There are totally six different types of sub-syllable units as listed in Table 3.1.

Unit type:	Patterns in synthesis unit:	Explanations:
Intra-syllable unit:	Initial–Final	A legitimate Cantonese syllable
Inter-syllable unit:	Final–Initial or Final–Final	The two units are from two adjacent syllables
Stand-alone units (a special type of inter-syllable units):	Silence–Initial or Silence–Final	The phoneme at the beginning of an utterance
	Final–Silence	The phoneme at the end of an utterance

Table 3.1 Different synthesis units in sub-syllable based system.

The following is an example of how to represent a Chinese sentence with sub-syllable units. Obviously, there is an overlapped Initial or Final unit between each pair of consecutive sub-syllable units. The intra-syllable and inter-syllable co-articulation are covered by the I-F and F-I units respectively. The concatenation between sub-syllable units is done within the overlapped segment. By doing so, all important transition regions can be retained [6].

Text input:	沙田 (Shatin - a place in Hong Kong)
Transcription:	/saa1-tin4/
Breaking down into sub-units:	/silence-s-aa1-t-in4-silence/
Sub-syllable synthesis unit list:	/silence-s/, /s-aa1/, /aa1-t/, /t-in4/, /in4-silence/

Table 3.2 An example in conversion from transcription to sub-syllable units.

3.3.2 Acoustic inventory

In Cantonese, there are about 1,800 tonal syllables. The acoustic inventory of the syllable-based system covers most of them. In the sub-syllable based system, the total number of sub-syllable units as defined in Table 3.1 is determined by the number of Initials, Finals and tone. In Cantonese, there are 19 Initials, 53 Finals (27 Finals may have null-Initial) and 6 tones. The total number of the possible combination of sub-syllable units is listed as follows:

Unit type	Total combination	Explanations
Initial-Final (Intra-syllable)	1,800	Same as the number of Cantonese tonal syllable
Final-Initial (Inter-syllable)	5,700	The calculation is based on: 19 Initials 300 tonal Finals 27 null-Initial Finals
Final-Final (Inter-syllable)	8,100	
Silence-Initial (stand-alone unit)	19	
Silence-Final (stand-alone unit)	27	
Final-Silence (stand-alone unit)	300	
TOTAL:	~16,000	

Table 3.3 Full set of sub-syllable units [6].

If the tonal difference is considered, the total number of distinctive sub-syllable units is about 16,000. For example, /s-aa1/ and /s-aa3/ are counted as different units, and /aa1-p/ and /aa2-p/ are counted as different units, because the Final units carry different tones. In [6], it was proposed to use a reduced set of sub-syllable units. Among the 16,000 sub-syllable units, many of them are differentiated only by tone. Since pitch modification would be applied, it becomes less critical to store all tonal variants of the Finals. In the reduced set of sub-syllable units, the six tones are divided into two broad classes, namely the level-tone group (Tone 1, 3, 4 and 6) and rising-tone group (Tone 2 and 5). For each Final, only one representative tone is

required to be included. As a result, the total number of sub-syllable units reduces to about 5,750. The distributions are shown as in Table 3.4.

As described in [6], each required sub-syllable unit was recorded as part of a carrier word, so that it is naturally spoken as far as possible. The lengths of the carrier words range from 2 to 5 characters. Each carrier word contains one designated sub-syllable unit. In the TTS system developed in [6], only such intended units are used for concatenative synthesis. However, each carrier word also contains a few sub-syllable units other than the intended unit. These unintended units may fall within the reduced set of the 5,750 sub-syllables. In this case, they can be used as an additional copy and made available for selection. Even if they are not within the required set, they may represent a useful tonal variation. In this study, we aim to fully exploit the recorded speech segments. It is found that there are 7,750 distinct sub-syllable units that can be used (see Table 3.4). Each of them may contain more than one copy. As a result, there are about 20,000 usable sub-syllable segments in the acoustic inventory.

Unit type	Reduced-set combination [6]	Total number of distinct units (different tones)
Initial-Final (Intra-syllable)	~1,000	~1,400
Final-Initial (Inter-syllable)	~1,900	~3,000
Final-Final (Inter-syllable)	~2,700	~3,000
Silence-Initial (stand-alone unit)	19	19
Silence-Final (stand-alone unit)	27	~70
Final-Silence (stand-alone unit)	~100	~250
TOTAL:	~5,750	~7,750

Table 3.4 Reduced set of sub-syllable units and total number of distinct units in the inventory.

The recorded words were segmented automatically by HMM forced alignment.

Marking of pitch peaks was done manually on all speech waveforms.

3.3.3 Determination of the concatenation points

In sub-syllable based system, waveform concatenation is done within the overlapped phonemic unit between a pair of sub-syllable units. This is different from the syllable based system in which concatenation is carried out at syllable boundaries. The concatenation strategies used in the TTS system described in [6] are briefly described below. Initials and Finals are divided into three different cases.

Case 1: Voiced units

All voiced units, including all Finals and voiced Initials (nasals, glides and semi-vowels) are handled in a similar way. The concatenation point is chosen to be at the beginning (the first 1/8 of the duration) of the speech segments, as the intensity level is relatively low.

Case 2: Plosives and affricates

Plosives and affricates are unvoiced units that typically contain a closure period. The concatenation point is located as the starting point of the speech segments, under the assumption this point of these segments should fall into the closure period.

Case 3: Fricatives

The concatenation point for a fricative Initial is located at the middle point of the fricative segment. That is, each of the sub-syllable unit contributes half of the synthesized fricative segment.

3.4 Problems

The primary motivation of using sub-syllable units for Cantonese is to retain the inter-syllable co-articulation [6]. Perceptual test showed that it produced more fluent synthetic speech than the syllable based system. Transitions between consecutive syllables were better captured by introducing the inter-syllable units. However, it is noticed that the sub-syllable based system does not perform well in many cases. The synthesized speech is definitely not considered to be highly natural. Sometimes it contained abrupt changes in speech signals and perceived artifacts. We believe that the problems are on the following two aspects:

1) Concatenation strategies:

   Each Cantonese phoneme has its own acoustic-phonetic properties. These properties should be considered in the design of concatenation method. In the existing system, the phonemes are roughly divided into three categories only. We believe that such groupings are too generalized. For example, there are several types of syllable structures in Finals. But the existing system does not further classify them into finer groups. Moreover, determination of the concatenation points does not consider the acoustic properties of the phonemes. Proposed strategies in existing system is too course and unsuitable.

2) Unit selection:

   In the existing system, all speech segments are selected from the intended units only. Previous calculations show that there are many unused units

inside the inventory, which is also usable in synthesizing speech. We should fully utilize all possible units which is stored inside the inventory. Since there may have more than one possibilities for each unit, so a selection process is required to choose the most appropriate speech segment.

In the coming chapters, we will have a detailed discussion in the concatenation strategies (chapter 4) and the unit selection process (chapter 5) of the proposed TTS system.

References:

[1] Tan Lee, Helen Meng, W. Lau, W. K. Lo & P. C. Ching, "Micro-prosodic control in Cantonese Text-to-Speech synthesis", in Proc. Eurospeech-99, Vol. 4, pp. 1855-1858, Budapest.

[2] E. Moulines & J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech", in Speech Communications Vol. 16, pp. 175 – 205.

[3] E. Moulines & F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", in Speech Communications Vol. 9, pp. 453 - 467.

[4] Shaw-Hwa Hwang, Sin-Horng Chen & Yih-Ru Wang, "A Mandarin text-to-speech system", in Proc. ICASSP-97, Vol. 3, pp. 1421 – 1424.

[5] M. Chu & P.C. Ching, "A Cantonese synthesizer based on TD-PSOLA method", in Proc. ISMIP-97, pp. 262-7, Taipei.

[6] K.M. Law, "Cantonese Text-to-Speech Synthesis Using Sub-syllable Units", M. Phil. Thesis, the Chinese University of Hong Kong, 2001

Chapter 4

Waveform Concatenation for Sub-syllable Units

Waveform concatenation refers to the process in which two given speech segments are joined together to form a new speech waveform. As the original segments are recorded separately with different phonetic contexts, the concatenation would inevitably lead to signal discontinuities or artifacts. For high-quality speech synthesis, it is important to minimize the perceptual effect of such artifacts. In this chapter, our goal is to develop a set of concatenation strategies for Cantonese sub-syllable units as described in Section 3.3.

4.1 Previous work in concatenation methods

Given a pair of speech segments, the concatenation typically involves two steps. Firstly, the concatenation point needs to be determined for each segment. Secondly, the two segments are joined at the concatenation points. In the simplest case, the two segments are joined end-to-end. The concatenation points are at the beginning and the ending of the segments. But more generally, the concatenation points lie in the middle of a segment. The signal samples in the neighborhood of the concatenation points need to be manipulated so as to minimize the undesirable discontinuities and artifacts. There have been studies on waveform concatenation techniques [1] [2]. Here we are going to provide a summary of them.

4.1.1 Determination of concatenation point

In general, the concatenation point is chosen to be located at: 1) syllable boundaries, or 2) stationary regions in which modification of the signal would be perceptually less sensitive.

1) Concatenation at syllable boundaries

A syllable has a general structure of consonant-vowel-consonant. If the concatenation is done at the syllable boundaries, only the consonant segments are involved. In this way, the entire vowel segment is preserved. This is desirable because vowels are the most prominent part in terms of signal intensity. Therefore if the concatenation is done in a vowel segment, it is more likely to introduce perceptually noticeable artifacts. In addition, if the concatenation is done at the syllable boundaries, the co-articulation from consonant to vowel can be largely retained.

2) Acoustically stationary regions

It is preferable to perform concatenation in a segment that is relatively stationary. As the signal spectrum and other acoustic properties change slowly, it would be easier to find out a concatenation point where the two segments can match well with each other. Phonetically, such stationary regions can be found in fricative, vowel and nasal segments.

4.1.2 Waveform concatenation

Speech segments can be joined either by hard concatenation or soft concatenation:

1) Hard concatenation

The speech segments are joined directly at the concatenation points. There is no modification to be made in the signals. Hard concatenation is usually applied only at the syllable boundaries. It is simple to implement.

2) Soft concatenation

Soft concatenation means that the signal in a close vicinity of the concatenation point is modified during the concatenation so as to make the transition as smooth as possible. For example, the transition region can be generated by "overlapped addition" of short windows from each of the segments.

4.2 Problems and difficulties in concatenating sub-syllable units

For the TTS system described in Chapter 3, the speech segments to be concatenated are a pair of sub-syllable units. Let U_1 and U_2 denote the first and the second sub-syllable segments respectively. They share the same phoneme p. As shown in Figure 4.1, $U_1 = [u_L][p_1]$ and $U_2 = [p_2][u_R]$, where p_1 and p_2 denote the acoustic realizations of p in the different contexts of U_1 and U_2 respectively. u_L and u_R are the neighbor phonemes of p.

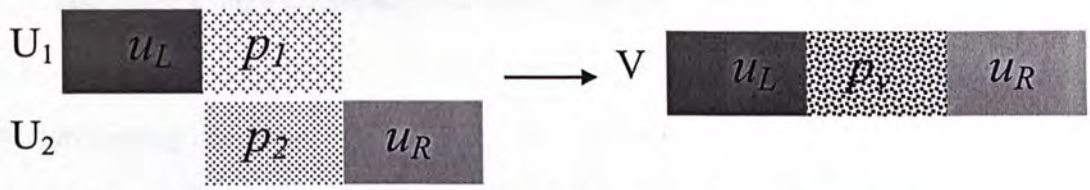


Figure 4.1 Graphical illustrations in concatenating two speech segments.

The result of concatenation can be denoted by $V = [u_L][p_v][u_R]$, where p_v is a new segment that carries the phoneme p. The major difficulty for the concatenation is due to the mismatch of acoustic properties between p_1 and p_2. For Cantonese, there also exists a problem on some of the Initials, which use the same phonemic symbol to represent a number of different acoustical realizations.

4.2.1 Mismatch of acoustic properties

Although p_1 and p_2 are carrying the same phonemes, they may be very different from each other in signal properties. The difference is due to either the linguistic contexts or the recording conditions, which may introduce large variations in spectral content, pitch level, duration, intensity, etc.

- *Linguistic contexts*

  Since Cantonese is a tonal language, p_v must carry not only the desired phoneme (Initial or Final) but also a specific tone. Ideally, p_1 and p_2 should carry exactly the same phoneme with the required tone. However, in the inventory of sub-syllable units as described in Section 3.3.2, it is not guaranteed that such a requirement can always be fulfilled. Instead, it is very often that p_1 and p_2 are in different tones. As a result, their pitch and harmonic structures would be very different. On the other hand, the co-articulation from neighboring phonemes may also cause a large discrepancy between p_1 and p_2.

- *Recording conditions*

  On the other hand, even if p_1 and p_2 are exactly the same in terms of linguistic contexts, they may also be acoustically different because of their different

recording conditions. Typically, p_1 and p_2 are excised from different carrier utterances, which were recorded at different time. There may be differences in the physical conditions of the speakers and the speaking rates. All of these factors lead to mismatch of acoustic properties between p_1 and p_2.

In Chapter 5, it will be seen that the unit selection process attempts to choose p_1 and p_2 that are acoustically as similar as possible. But they will never be exactly the same. Sometimes the deviations among all available segments are very large. For example, the pitch level of all segments containing the Final /aam/ with Tone 1 varies from 202 Hz to 266 Hz. There may be cases that the best selected pair of p_1 and p_2 has very different pitch levels.

Supposedly we are not going to change the acoustic properties of p_1 and p_2 substantially during the concatenation. However, if they are too different, applying some modifications on the signals become inevitable. This may affect the quality of the output speech. An alternative way is to simply abandon either one of p_1 and p_2, i.e. let $p_v = p_1$ or $p_v = p_2$. The consequence is that the co-articulation effect carried by the abandoned segment would be lost. In fact, retaining the co-articulation effect is part of the motivation why we use sub-syllable based units.

The concatenation strategies and the unit selection process discussed in the current and next chapters are developed from the acoustical point of view. We attempt to minimize the mismatch and the modification of acoustic properties during the synthesis process.

4.2.2 Allophone problem of Initials /z/, /c/ and /s/

An allophone is a phonetic variant of a phoneme in a particular language [3] [4]. In Cantonese, each of the Initials /z/, /c/ and /s/ has different phonetic realizations, as shown in Table 4.1. In spoken Cantonese, people would not pronounce it wrongly because the following Final uniquely determines the choice of the exact realization. Therefore, in the LSHK transcription scheme, the same symbol is used to label different pronunciations, although they are represented by different IPA symbols.

LSHK symbol	IPA symbol	Example			
		LSHK symbol	IPA symbol	Chinese character	English meaning
z	ts	zi1	tsi1	知	know
	tʃ	zoeng1	tʃœŋ1	將	will
c	t^hs	caa4	t^hsa4	查	examine
	t^hʃ	cyu5	t^hʃy5	儲	store
s	s	saa1	sa1	沙	sand
	ʃ	syut3	ʃyt3	雪	snow

Table 4.1 Different phonetic realization of the Initials /z/, /c/ and /s/

For the concatenation of sub-syllable units $U_1 = [u_L][p_1]$ and $U_2 = [p_2][u_R]$, if the overlapped phoneme p is one of such allophonic units, there is a possibility that p_1 is not the desired allophonic variation. This is because the right context from which U_1 is extracted may not be the same as u_R. For example, consider the concatenation of the sub-syllable units /un1-s/ and /s-yu1/. If the original right context of the segment /un1-s/ is the Final /aa/, then the two /s/ segments to be concatenated would differ greatly. This is illustrated by the example as shown in Figures 4.2 and 4.3.

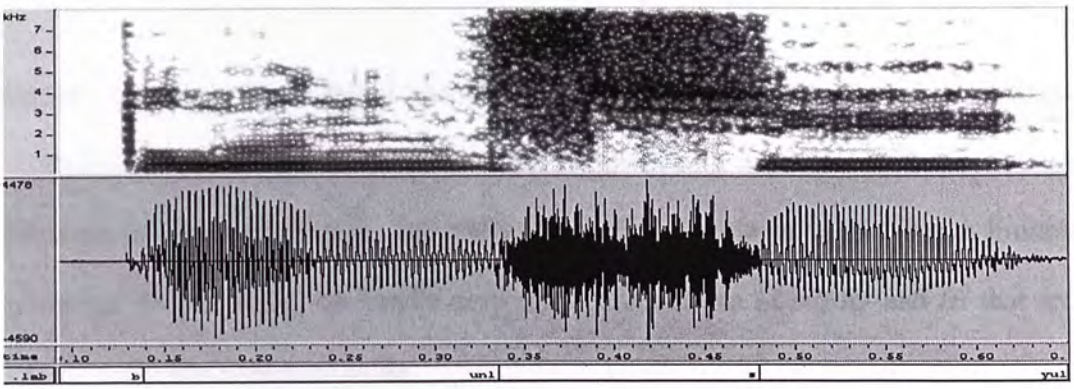


Figure 4.2 Concatenated speech of /bun1-syu1/ (搬書) showing there is a spectral mismatch in phoneme /s/.

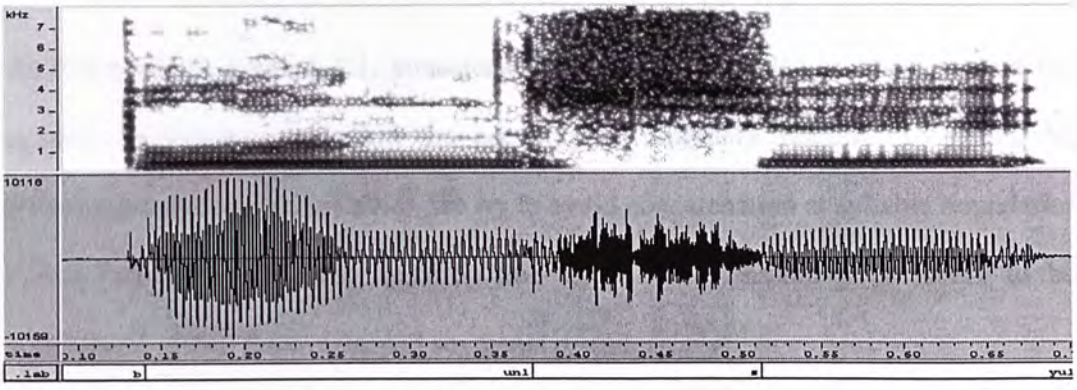


Figure 4.3 Concatenated speech of /bun1-syu1/ (搬書) after implementing the allophonic variations.

To avoid such mismatch of acoustic properties, we divide the vowel nucleus into two groups, which correspond to two most distinctive allophonic variations of /z/, /c/ and /s/. As shown in Table 4.2, two different symbols are used to represent these variations. In this way, for the concatenation of these Initials, the selected units would be better matched in their phonetic and acoustic properties.

Initials (Allophonic units only)			Vowel nucleus
z	c	s	a, aa, e, eo, i, o, u
zh	ch	sh	oe, yu

Table 4.2 Two different symbols are used in representing allophonic units.

Strictly speaking, other Initial and Final segments may also have such allophonic variations that are determined by the context. However, it is practically impossible to label all of them differently, especially when the acoustic inventory has a limited coverage. In our study, we handle only the two variations of /z/, /c/ and /s/ that are perceptually the most noticeable.

4.3 General procedures in concatenation strategies

As discussed in Section 4.1, concatenation points are usually assigned around the syllable boundaries or within the acoustically stationary regions. To retain the co-articulation between syllables, we try to avoid concatenation at syllable boundaries. For a pair of sub-syllable units U_1 and U_2, a concatenation point needs to be determined within the overlapped segments p_1 and p_2.

Each phoneme has its own acoustic and phonetic properties, which may be similar or very different with other phonemes. We need to consider these properties, especially from the acoustical point of view, as much as possible. Generally, concatenation may be done in either an unvoiced speech segment or a voiced speech segment.

For an unvoiced consonant segment, we avoid doing concatenation in the transitory region as far as possible. In Cantonese, plosive and affricate Initials typically contain a closure period in which the signal intensity is very low. Such closure region is a good place for concatenation of speech segments. On the other hand, fricative segments are acoustically stationary without periodicity contents. Theoretically concatenation can take place anywhere within a fricative segment.

For a voiced segment, we need to find a pair of corresponding points in p_1 and p_2, which have the best-matched spectra. By spectral match, we refer mainly to the match of formant peaks.

4.3.1 Concatenation of unvoiced segments

Unvoiced speech segments are jointed by simple hard concatenation. That is, the two segments are connected directly without any modification on the waveforms. Hard concatenation can be used for the concatenation of fricatives and other unvoiced segments that are acoustically stationary. It can also be used for plosives and stops that contain a closure region.

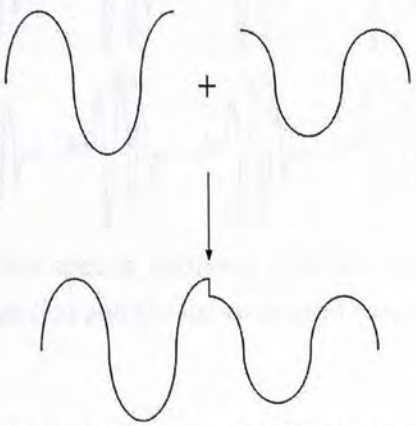


Figure 4.4 Graphical illustrations in concatenating two waveforms by hard concatenation.

4.3.2 Concatenation of voiced segments

Voiced speech segments are short-time stationary with clear periodicity in time domain. The speech waveform is composed of a sequence of periodic pitch cycles. In

each pitch cycle, the point with locally maximum amplitude is named pitch peak. Spectral analysis is done for each of the pitch cycles. A spectral distance can be computed between any pairs of pitch cycles. The concatenation point is then determined to attain the best spectral match, i.e. the smallest spectral distance. Soft concatenation technique is used to provide a smooth transition around the concatenation region. Detailed procedures are described as follows.

1) Identify all available pitch peaks within each speech segment. Pitch peaks that are near the syllable boundaries are ignored. In our implementation, about 10% of the pitch peaks (5% from each end) are not used.

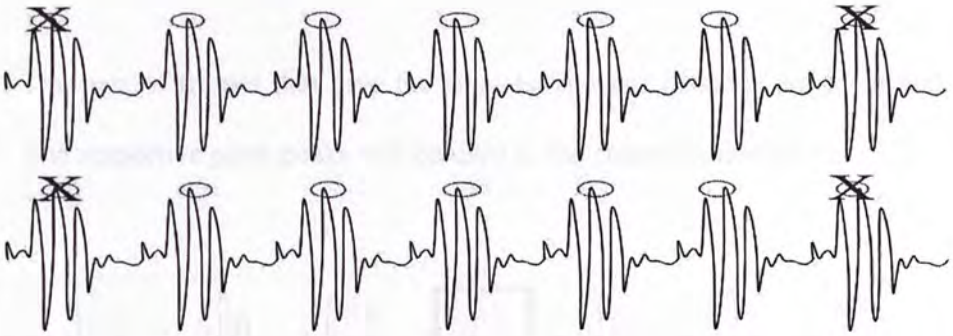


Figure 4.5 Two speech segments with the indication of the pitch peaks (circle) and the unconsidered cases (mark with X).

2) Each speech segment is divided into pitch-synchronous short-time frames. They are multiplied by a sequence of pitch-synchronous Hamming windows. These windows are centered at the pitch peak location. Afterwards, the short-time spectrum of each frame is calculated.

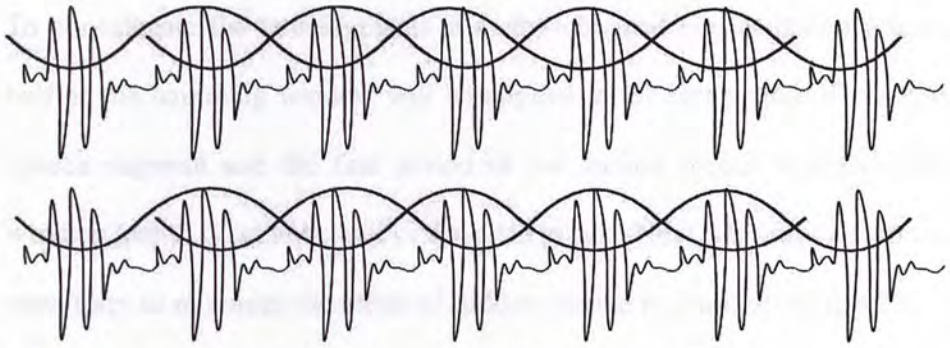


Figure 4.6 Speech segments applying Hamming windows.

3) Let denote the two speech segments to be concatenated as A and B. For each frame in segment A, its distance with each frame in segment B is computed. For different phonemes, we use different spectral distance measurements. Details will be given in the next section.

4) The pair of frames that give the smallest spectral distance are identified. The respective pitch peaks will be used as the concatenation points.

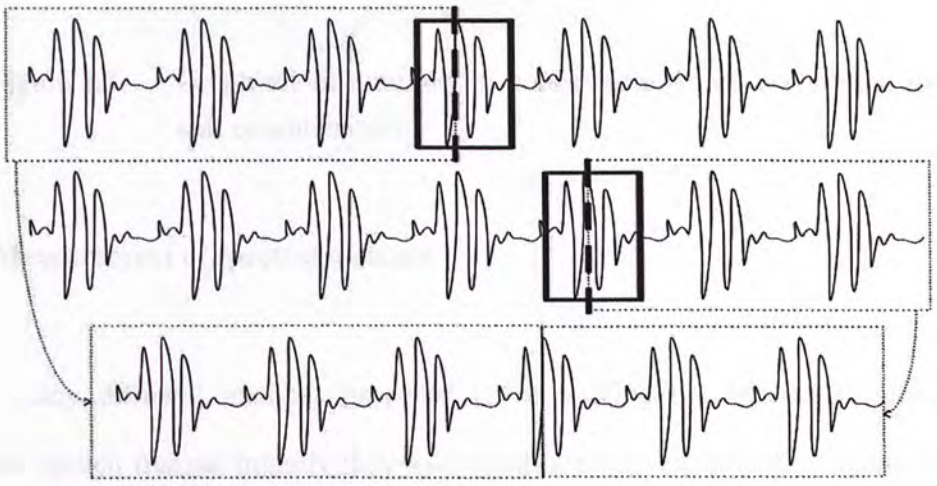


Figure 4.7 Selected window-pair with smallest spectral difference (solid square window) and concatenation of two speech segments (dotted windows)

5) To concatenate the two segments with the obtained concatenation points, half of the hamming window will be applied to the last period of the first speech segment and the first period of the second speech segment. The window length is set to be half of the pitch period. Next, we will overlap the waveform to minimize the effect of sudden change in synthesized speech.

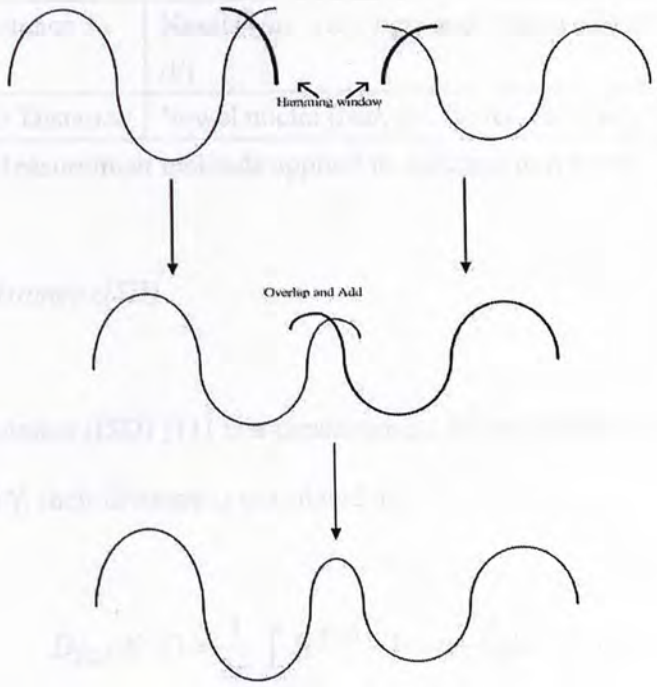


Figure 4.8 Graphical illustrations in concatenating two waveforms by soft concatenation.

4.3.3 Measurement of spectral distance

There are many different methods proposed for measuring the spectral distance between two speech frames. Initially they were used in automatic speech recognition [5]. The methods included log spectral distance, Itakura-Saito distance, Kullback-Leibler distance and so on. In concatenative speech synthesis, researchers have proposed different spectral distance measurements [6] [7] [8]. The work by J.

Wouters and M. W. Macon found that Itakura-Saito distance performed better in perceptual evaluation of voiced units [9]. Stylianou and Syrdal showed that Kullback-Leibler distance with FFT spectra performed better in concatenating vowel nuclei [10]. In our study, these two measurements are used. Their usage for Cantonese phonemes is shown as in the following table.

Itakura-Saito Distance	Nasals (/m/, /n/, /ng/) and other voiced Initials (/j/, /w/, /l/)
Kullback-Leibler Distance	Vowel nuclei (/aa/, /e/, /i/, /o/, /u/, /oe/, /yu/, /a/, /eo/)

Table 4.3 Measurement methods applied in different unit types.

Itakura-Saito Distance (ISD)

Itakura-Saito distance (ISD) [11] is a measurement of log likelihood ratio. Given two spectrum X and Y, their distance is calculated as:

$$D_{ISD}(X,Y) = \frac{1}{2\pi}\int_{-\pi}^{\pi}[e^{V(\omega)} - V(\omega) - 1]d\omega \quad \ldots (1)$$

$$\text{where } V(\omega) = \log\frac{X(\omega)}{Y(\omega)} \quad \ldots (2)$$

This measurement is asymmetric. It can be modified to give the symmetric Itakura-Saito distance as follows:

$$D_{SISD}(X,Y) = \frac{1}{4\pi}\int_{-\pi}^{\pi}[\frac{X(\omega)}{Y(\omega)} + \frac{Y(\omega)}{X(\omega)} - 2]d\omega \quad \ldots (3)$$

Kullback-Leibler Distance (KLD)

Kullback-Leibler distance (KLD) [12] is to measure the "distance" between two probabilities. Consider there are two feature vectors X and Y, its distance will be defined as:

$$D_{KLD}(X,Y)=\frac{1}{2\pi}\int_{-\pi}^{\pi}X(\omega)\log\frac{X(\omega)}{Y(\omega)}d\omega \quad \ldots (4)$$

This measurement is asymmetric. It can be modified to give the symmetric Kullback-Leibler distance as follows:

$$D_{SKLD}(X,Y)=\frac{1}{4\pi}\int_{-\pi}^{\pi}[X(\omega)-Y(\omega)]\log\frac{X(\omega)}{Y(\omega)}d\omega \quad \ldots (5)$$

4.4 Detailed procedures in concatenation points determination

Unvoiced and voiced speech units are treated separately. Units with similar acoustic properties are grouped into the same class, to which the same concatenation strategy is used to fit these properties.

4.4.1 Unvoiced segments

Unvoiced segments include some of the Initials, i.e., plosives, affricates and fricatives. They are of transitory nature. The acoustic properties change rapidly across the segments. It is difficult to find a truly stable point within such a segment. Based on their manner of articulation and transitory properties, the unvoiced Initial units are

divided into two main categories as shown in Table 4.4.

Groupings	Examples
Plosive and Affricate	(Plosives) /b/, /d/, /g/, /p/, /t/, /k/, /gw/, /kw/ (Affricates) /z/, /zh/, /c/, /ch/
Fricative	/f/, /s/, /sh/, /h/

Table 4.4 Groupings of unvoiced segments into different classes

Plosives and affricates

Plosives and affricates are burst-like sound. Voice is produced by stopping the airflow in the vocal tract and to be followed by a sudden release in producing such sounds. The amplitude of the signal is small when the airflow is blocked. Concatenation is done preferably in such a closure period. It is believed that no significant audible artifact would be generated.

For all plosive and affricate speech segments, we experimentally measure the length of their closure period. It is defined as the period during which the sample amplitudes are always below a certain threshold. Although theoretically there should exist a closure period in these segments, sometimes it can be very short, due to various reasons. For example, there are occasionally short spikes caused by lip smacks or other unintended articulatory movement. If a closure period is too short, perceptually people cannot detect such region clearly and unable to recognize the plosives or affricates due to the lack of this special property. So a reasonable length of the closure period should be able to detect for all plosives and affricates.

Table 4.5 gives the statistics of short closures in the acoustic inventory of sub-syllable

units. A short closure is defined as one shorter than 0.01 second. Two different threshold values (200 and 400) are tested for 16-bit speech samples.

Plosive / Affricate	Available in inventory	Mean duration of closure (second) (threshold: 200)	No. of units with short closure (threshold: 200)	Percentages	Mean duration of closure (second) (threshold: 400)	No. of units with short closure (threshold: 400)	Percentages
b	494	0.056	12	2.43%	0.059	10	2.02%
d	492	0.044	18	3.66%	0.047	7	1.42%
g	934	0.042	33	3.53%	0.044	18	1.93%
p	190	0.036	15	7.89%	0.044	5	2.63%
t	333	0.039	32	9.61%	0.044	13	3.90%
k	346	0.038	31	8.96%	0.042	13	3.76%
gw	185	0.027	33	17.84%	0.046	25	13.51%
kw	125	0.025	29	23.20%	0.034	10	8.00%
z / zh	1077	0.033	89	8.26%	0.036	40	3.71%
c / ch	500	0.021	128	25.60%	0.025	90	18.00%

Table 4.5 Statistics on the closure period of the plosive and affricate units in our speech database.

From the experiment, we found that most of the plosives and affricates contain a clear closure period ranging from 0.03 to 0.05 seconds. However, we note that some phonemes may need a higher threshold value in order to have a better detection of the closure. We suggest using 200 as the threshold value in most of the cases. For the phonemes /gw/, /kw/, /c/ and /ch/, a higher threshold of 400 is used. Concatenation points are determined at the start points of both closure regions.

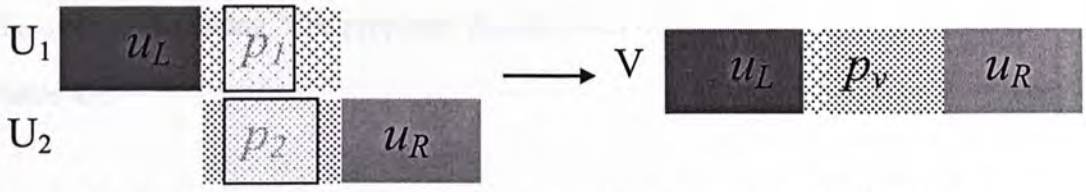


Figure 4.9 Graphical illustrations in concatenating plosives and affricates. (p_1, p_2 and p_v: Initials; u_L and u_R: Finals; white color: closure region)

Fricatives

Fricative sounds are produced by the friction in the narrow opening of the vocal tube. The segment has fairly stationary spectral property. Typically, it has high-energy concentration at high frequency band. Waveform concatenation can be done at any point within the segment, except for the segment boundaries (either beginning or ending region) of the speech segments, which are less stationary because of the co-articulation effects. Our strategy is to apply concatenation at the middle of the speech segments. This is the same method as used in the existing system.

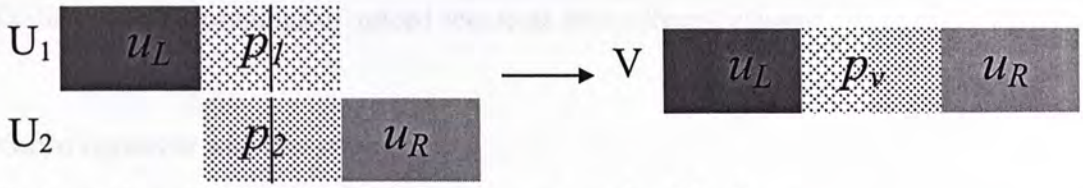


Figure 4.10 Graphical illustrations in concatenating fricatives.
(p_1, p_2 and p_v: Initials; u_L and u_R: Finals; black lines: middle point of phoneme p)

4.4.2 Voiced Segments

Voiced segments include some of the voiced Initials, i.e. nasals, glides and liquids, and all Finals. A Final unit is made up of a vowel nucleus and an optional coda. The coda may be either a nasal (/m/, /n/, /ng/), a stop (/p/, /t/, /k/) or a vowel (/i/ and /u/) in the case of diphthong. The proposed classification of voiced segments is shown as in Table 4.6.

No. of phones in the segment	Group	Description of phones	Examples
1	N_I	nasal Initials	/m/, /n/, /ng/
	O_I	other voiced Initials	/j/, /w/, /l/
	V	vowel nucleus only	/aa/, /e/, /i/, /o/, /oe/, /u/, /yu/
	N	syllabic nasal	/m/, /ng/
2	D	vowel nucleus + vowel coda	/aai/, /aau/, /iu/, /ui/, /eoi/, /oi/, /ai/, /au/, /ei/, /ou/
	VN	vowel nucleus + nasal coda	/aam/, /aan/, /aang/, /am/, /an/, /ang/, /eng/, /eon/, /im/, /in/, /ing/, /on/, /ong/, /oeng/, /un/, /ung/, /yun/
	VS	vowel nucleus + stop coda	/aap/, /aat/, /aak/, /ap/, /at/, /ak/, /ek/, /eot/, /ip/, /it/, /ik/, /ok/, /ot/, /oek/, /ut/, /uk/, /yut/

Table 4.6 Groupings of voiced segments into different classes

Voiced segments with one phone

There are 15 voiced segments that contain only one phone. They are the "N_I" group, "O_I" group, "V" group and "N" group. These segments have periodic waveforms. If they are sufficiently long, we may be able to find a period with relatively stationary spectral properties. The method discussed in Section 4.3.2 will be applied.

Voiced segments with two phones

The segments with two phones include the "D" and the "VN" group. There exists a transition between the two components. When determining the concatenation points, we need to ensure that the concatenation points from the two units are from the same phonetic component. For example, *p1* and *p2* are in the same phoneme /aam/. The concatenation points for both *p1* and *p2* should be determined either in the region of

the first component (i.e. roughly corresponds to /aa/) or the second component (i.e. roughly corresponds to /m/). For each of the units in Table 4.6, we select one of the phonetic components and restrict the concatenation point within this component.

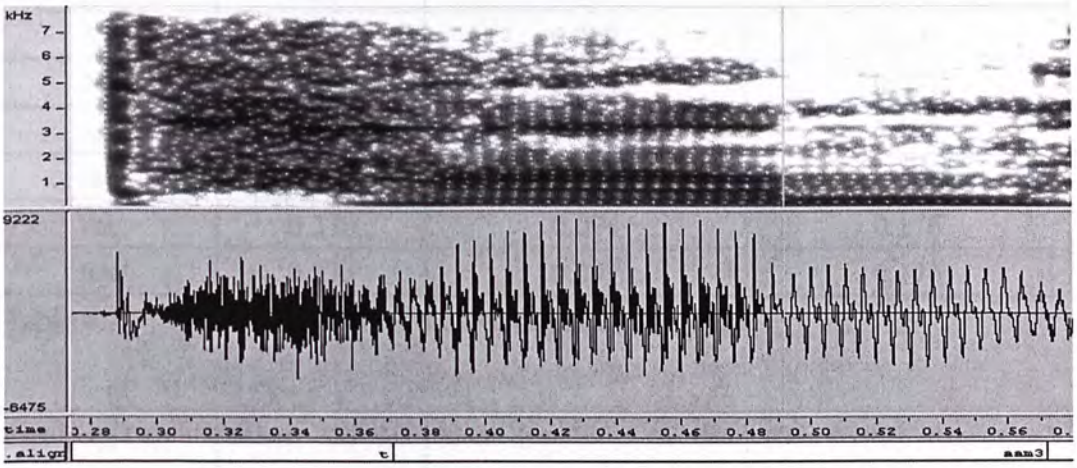


Figure 4.11 Transition point between two phonemes are marked by a straight line in the example of Chinese character /taam3/ (探).

We believe that a phone with relatively long duration would be more likely to contain a stationary region, in which a more reliable concatenation point can be determined. Therefore the mean duration for each component within these speech segments has been analyzed. For all speech segments in the "D" and "VN" group, we manually mark the transition point between the two phonetic components and measure the duration of these two components. The statistical mean values are given in the following tables.

Diphthong Unit	Mean duration (second)	Mean duration of the first component (second)	Mean duration of the second component (second)
/aai/	0.329	0.188	0.141
/aau/	0.352	0.207	0.145
/eoi/	0.246	0.147	0.099
/iu/	0.218	0.139	0.080
/oi/	0.308	0.200	0.108
/ui/	0.195	0.124	0.071
/ai/	0.316	0.124	0.192
/au/	0.266	0.129	0.136
/ei/	0.270	0.094	0.176
/ou/	0.273	0.097	0.176

Table 4.7 Mean duration of the two regions in D group

Diphthong Unit	Mean duration (second)	Mean duration of the first component (second)	Mean duration of the second component (second)
/aam/	0.323	0.187	0.136
/aan/	0.318	0.177	0.141
/aang/	0.367	0.199	0.168
/eng/	0.266	0.135	0.131
/on/	0.318	0.189	0.128
/ong/	0.311	0.174	0.137
/oeng/	0.229	0.119	0.110
/yun/	0.218	0.110	0.107
/am/	0.295	0.103	0.192
/an/	0.262	0.100	0.162
/ang/	0.312	0.123	0.189
/eon/	0.230	0.083	0.147
/im/	0.211	0.089	0.122
/in/	0.227	0.111	0.116
/ing/	0.224	0.085	0.139
/un/	0.191	0.093	0.098
/ung/	0.258	0.113	0.145

Table 4.8 Mean duration of the two regions in VN group

Based on the statistics, we divide the units in the "D" and "VN" groups into two

categories, from which the concatenation point will be searched in the first and the second phonetic components respectively. The remaining procedures, including the determination of concatenation points and the concatenation methods, are similar to the case of voiced segments with one phone.

Concatenation point to be searched in the first component	(D group) /aai/, /aau/, /eoi/, /iu/, /oi/, /ui/ (VN group) /aam/, /aan/, /aang/, /eng/, /on/, /ong/, /oeng/, /yun/
Concatenation point to be searched in the second component	(D group) /ai/, /au/, /ei/, /ou/ (VN group) /am/, /an/, /ang/, /eon/, /im/, /in/, /ing/, /un/, /ung/

Table 4.9 Further classifications for voiced segments with two phonemic components.

VS: vowel with stop coda

A segment in this group ends with a stop coda, either /p/, /t/ or /k/. These transient phones are not released. A closure period is found at the end of the segment due to the sudden stop of the airflow in articulation. Similar to the case of plosives and affricates in Initials, concatenation is done preferably in such a closure region. The concatenation point for a VS unit is determined at the end points of the detected closure period. In this case, we will follow the procedures as for the plosive and affricate Initials.

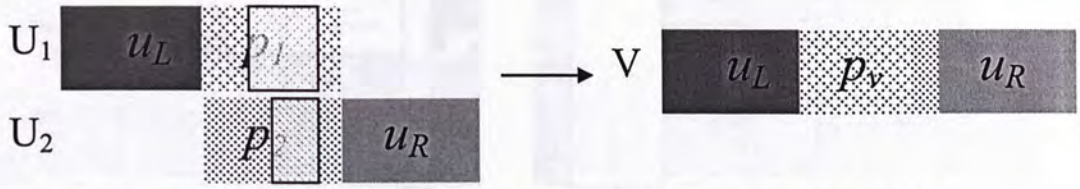


Figure 4.12 Graphical illustrations in concatenating vowel with stop coda. (p_1, p_2 and p_v: Finals; u_L and u_R: Initials and/or Finals; white color: closure region)

4.5 Selected examples in concatenation strategies

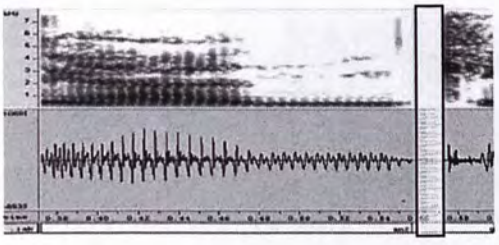
In this section, we will show a number of typical examples of concatenating different kinds of sub-syllable units in our TTS system. Both the signal waveforms and spectrograms will be shown for illustrating the procedures and results.

4.5.1 Concatenation at Initial segments

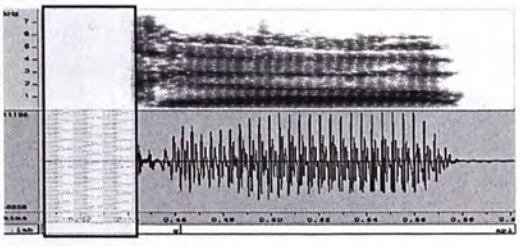
In this case, we need to concatenate an inter-syllable unit (Final-Initial or Silence-Initial) and an intra-syllable unit (Initial-Final). Two cases will be studied: a plosive and a fricative.

4.5.1.1 Plosives

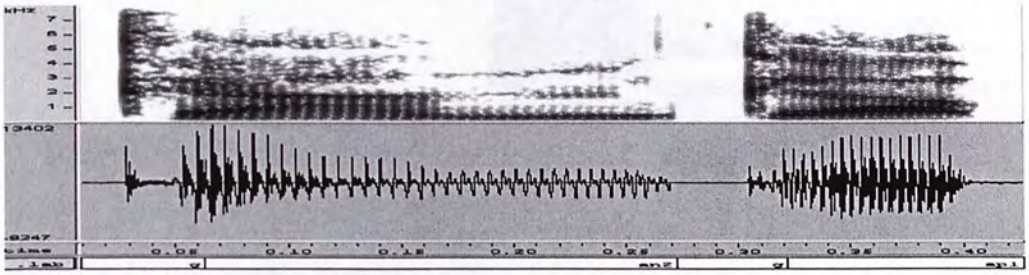
Figure 4.13 shows an example of concatenating the unit /an2-g/ and /g-ap1/ to form the word /gan2-gap1/, meaning "urgent" in English. A closure period is seen within the two speech segments and concatenation is done there.



(a) Speech segment containing the inter-syllable /an2-g/



(b) Speech segment containing the intra-syllable /g-ap1/

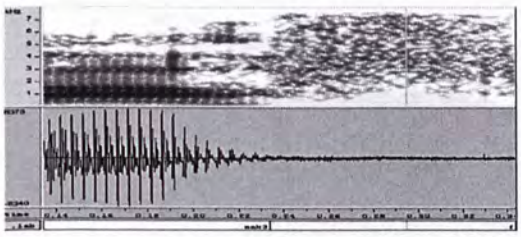


(c) Synthesized speech segment /gan2-gap1/

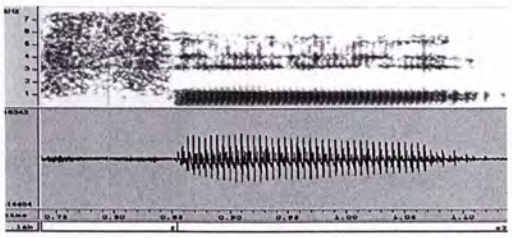
Figure 4.13 Example of concatenating plosive.

4.5.1.2 Fricatives

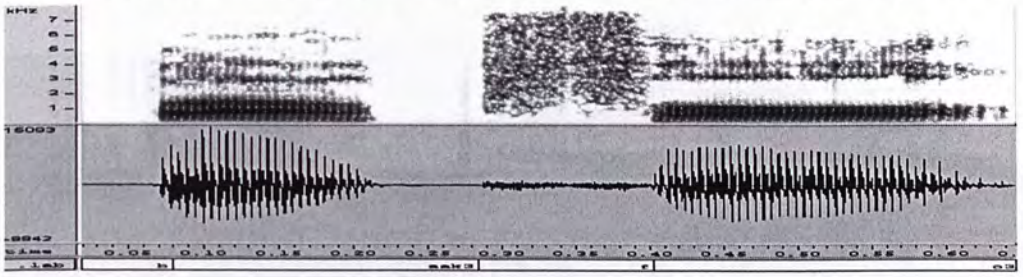
Figure 4.14 shows an example of concatenating the unit /aak3-f/ and /f-o3/ to form the word /baak3-fo3/, meaning "department store" in English. Concatenation points are determined in the middle of the speech segments.



(a) Speech segment containing the inter-syllable /aak3-f/



(b) Speech segment containing the intra-syllable /f-o3/



(c) Synthesized speech segment /baak3-fo3/

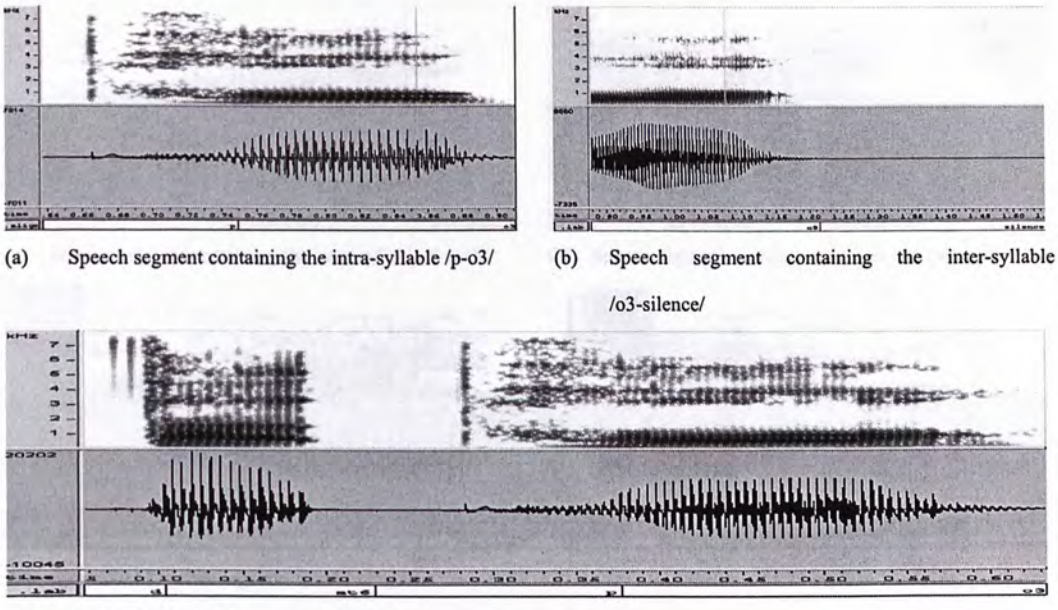
Figure 4.14 Example of concatenating fricative.

4.5.2 Concatenation at Final segments

An intra-syllable (Initial-Final) or an inter-syllable unit (Final-Final or Silence-Final) is concatenated with an inter-syllable unit (Final-Initial, Final-Final or Final-Silence). The following examples show the concatenation for a long vowel and a diphthong respectively.

4.5.2.1 V group (long vowel)

Figure 4.15 shows an example of concatenating the unit /p-o3/ and /o3-silence/ to form the word /dat6-po3/, meaning "breakthrough" in English. We apply spectral distance measurements to find out the most suitable point for concatenation. Kullback-Leibler distance will be applied in the mentioned calculations.



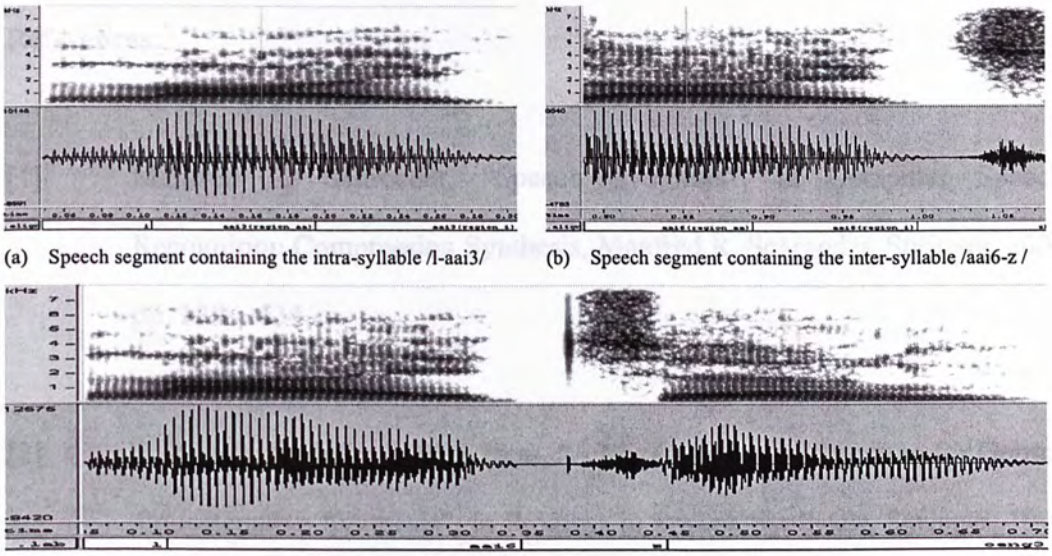
(a) Speech segment containing the intra-syllable /p-o3/ (b) Speech segment containing the inter-syllable /o3-silence/

(c) Synthesized speech segment /dat6-po3/

Figure 4.15 Example of concatenating long vowel.

4.5.2.2 D group (diphthong)

Figure 4.16 shows an example of concatenating the unit /l-aai3/ and /aai6-z/ to form the word /laai6-zoeng3/, meaning "refuse to pay a debt" in English. As discussed in Section 4.4.2, the Final /aai/ consists of two phonetic components. The concatenation is done in the first component, which roughly corresponds to /aa/. Kullback-Leibler distance is applied in calculating the spectral distance measurements of region /aa/.



(a) Speech segment containing the intra-syllable /l-aai3/ (b) Speech segment containing the inter-syllable /aai6-z /

(c) Synthesized speech segment /laai6-zoeng3/

Figure 4.16 Example of concatenating diphthong.

References:

[1] Manfred R. Schroeder, "Speech Synthesis", in Computer Speech: Recognition Compression Synthesis, Manfred R. Schroeder, Springer, 2004, pp. 129 – 134.

[2] T. Portele, F. Hofer & W. J. Hess, "A Mixed Inventory Structure for German Concatenative Synthesis", in Progress in Speech Synthesis, Springer, 1996, pp. 263 – 277.

[3] Eugene E. Loos, Susan Anderson, Dwight H., Day, Jr., Paul C. Jordan & J. Douglas Wingate, "Glossary of linguistic terms", SIL International, http://www.sil.org/linguistics/glossaryoflinguisticterms/contents.htm, Internet version, 2004.

[4] "Allophone", http://www.encyclopedian.com/al/Allophone.html, Encyclopedian (Online dictionary and encyclopedia), Internet version.

[5] L. Rabiner & B. H. Juang, "Fundamental of Speech Recognition", Prentice Hall Inc, New Jersey, 1993.

[6] S. Narayanan & A. Alwan, "Text-to-Speech Synthesis: New Paradigms and Advances", New Jersey: Pearson Education, Inc, 2005, pp 42-46

[7] Robert E. Donovan, "A New Distance Measure for Costing Spectral Discontinuities in Concatenative Speech Synthesisers", in 4th ISCA Tutorial

and Research Workshop on Speech Synthesis, Pethshire, Scotland, 2001, pp 59-62

[8] E. Klabbers & R. Veldhuis, "Reducing Audible Spectral Discontinuities", in IEEE Transactions on Speech and Audio Processing, Vol. 9 no. 1, 2001, pp 39-51

[9] J. Wouters & M. W. Macon, "A Perceptual Evaluation of Distance Measures for Concatenative Speech Synthesis", in Proc. ICSLP vol. 6, Sydney, Australia, 1998, pp 2747-2750

[10] Y. Stylianou & A.K. Syrdal, "Perceptual and Objective Detection of Discontinuities in Concatenative Speech Synthesis", in Proc. ICASSP, Vol. 2, 2001, pp 837-840

[11] F. Itakura & S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies", Electronics and Communications, vol. 53-A, Japan, 1970, pp 36-43

[12] S. Kullback & R.A. Leibler, "On Information and Sufficiency", in the Annals of Mathematical Statistics, Vol. 22 no. 1, 1951, pp 79-86

Chapter 5

Unit Selection for Sub-syllable Units

Unit selection in corpus-based TTS system refers to the process in which the best speech segment among the availability inside the inventory is selected. A speech utterance is synthesized with speech segments that come from different phonetic and acoustic contexts. The difference among the acoustic parameters of these segments may be very large, leading to undesirable mismatch and discontinuities at the concatenation points. If the inventory contains multiple copies of a speech unit, a selection process can be applied to find the most suitable one among the candidates. In this chapter, we will describe the selection process for the Cantonese sub-syllable units.

5.1 Basic requirements in unit selection process

5.1.1 Availability of multiple copies of sub-syllable units

As described in Chapter 3, there are over 20,000 sub-syllable units available in the acoustic inventory. The number of distinct tonal units is about 7,750. For a particular sub-syllable unit, there are possibly multiple speech segments that are available for our choice. If we ignore the tonal difference, the number of distinct units is even smaller and there would be more choices for each unit. In the following sections, we are going to define several different levels of "identical", which will form the basis of the subsequent discussion on unit selection.

5.1.1.1 Levels of "identical"

Consider a sub-syllable unit S = [$pa(ta)$][$pb(tb)$], where pa and pb denote the two constituent phonemes (Initial or Final). If pa and pb are Finals, then ta and tb denote the respective tones associated with them. We use three levels in describing the status of "identical" among the sub-syllable units.

Given a target sub-syllable segment S_t = [$pa_t(ta_t)$][$pb_t(tb_t)$], at least one segment S_r = [$pa_r(ta_r)$][$pb_r(tb_r)$] can be found, such that $pa_t = pa_r$ and $pb_t = pb_r$. This is the basic requirement for the design of the acoustic inventory (see Chapter 3). In this case, S_r does not have to carry the same tone as S_t. This is referred to as level 3 identical. For example, S_1 = /b-aa1/, S_2 = /b-aa2/, S_3 = /b-aa3/ and S_4 = /b-aa4/ are all in level 3 identical.

The definition of level 1 and level 2 identicals requires certain degree of similarities in ta and tb. It was discussed that the six tones of Cantonese can be divided into two tone classes: level-tone class (Tone 1, 3, 4 and 6) and rising-tone class (Tone 2 and 5). This classification was used in the design of our sub-syllable unit inventory [1]. Within the same tone class, the prosodic difference will be smaller than the case between different tone classes. It is expected that more modification would be required to modify a rising pitch contour to a flat contour and vice versa. Level 2 identical is defined as the tones associated with the sub-syllable units that come from the same tone class. For the same example as above, S_1, S_3 and S_4 are regarded as being level 2 identical. All of them belong to the level-tone class.

Level 1 identical uses the most restrictive definition. It is required that the units being

compared have exactly the same tone identities, i.e., $ta_t = ta_r$ and $tb_t = tb_r$. For example, given S_1 = /b-aa1/, S_2 = /b-aa4/ and S_3 = /b-aa1/, S_1 and S_3 belong to the same in level 1 identical.

According to the above definitions, level 1 identical is the finest while level 3 identical is the most general one. A finer identical level generally means a better match in prosodic parameters. For unit selection, we can start from the finest level. If no suitable candidates can be found, we may relax to other levels to include more units in the selection process.

5.1.1.2 Statistics on the availability

In this section, we will provide the statistics on the availability of units according to the three levels of identical described above. If a sub-syllable unit has more than one copy available for selection based on a particular identical level, it is said to be a *multi-copy unit*. Otherwise, it is called a *unique-copy unit*. Whether a particular sub-syllable unit is a multi-copy unit or a unique-copy unit depends on how we define "identical", i.e. level 1 to 3.

Level 1 identical

There are about 7,750 distinct units based on the level 1 identical definition. Table 5.1 summarizes the availability of different types of units. There are 22% of the units having multiple copies. With such a restrictive definition, there would be very little room for the selection process to be effective. Because the concatenated speech segments are subject to prosodic modification, it would be reasonable to consider

other units that carry different tones. In fact, the exact pitch contour of a tone may vary greatly in real speech. Units with different tones do not necessarily have very different pitch values.

Combinations	No. of distinct units	No. of multi-copy units	Percentage of multi-copy units
Silence-Initial	22	22	100.0%
Silence-Final	71	40	56.3%
Initial-Final	1395	809	58.0%
Final-Initial	2981	614	20.6%
Final-Final	3028	1	0.03%
Final-Silence	252	232	92.1%
Overall	7749	1718	22.2%

Table 5.1 Statistics in multi-copy units using level 1 identical definition.

Level 2 identical

As shown in Table 5.2, about one third of the distinct units have multiple copies at this level. Recall that Cantonese syllable has the Initial-Final structure. Although null Initial is allowed, its frequency of occurrences in Cantonese speech is much lower than the Initial-Final combinations. If we do not consider the cases involving null Initial, the percentage of multi-copy units is about 49%. That means, nearly half of the distinct units contains multiple copies.

Combinations	No. of distinct units	No. of multi-copy units	Percentage of multi-copy units
Silence-Initial	22	22	100.0%
Silence-Final (flat)	49	41	83.7%
Silence-Final (rising)	8	2	25.0%
Initial-Final (flat)	610	442	72.5%
Initial-Final (rising)	368	193	52.5%
Final (flat)-Initial	1146	572	49.9%
Final (rising)-Initial	906	206	22.7%
Final (flat)-Silence	53	53	100.0%
Final (rising)-Silence	46	45	97.8%
Final (flat)-Final (flat)	1325	300	22.6%
Final (flat)-Final (rising)	108	25	23.2%
Final (rising)-Final (flat)	1150	0	0.0%
Final (rising)-Final (rising)	93	0	0.0%
Overall	5884	1901	32.3%
Overall (segments with null Initial not considered)	3151	1533	48.7%

Table 5.2 Statistics in multi-copy units using level 2 identical definition.

Level 3 identical

With this most general identical level, the total number of distinct units is only 3335. Statistics in multi-copy units using definition in level 3 identical are summarized as in Table 5.3. In this case, more than 80% of the sub-syllable units contain multiple copies. This provides us a lot of choices when selecting units. Yet, all tone classes are mixed together. It is more likely to come across a unit with an unmatched tone. Acoustic variations between different tones may be so large that substantial pitch modification is needed.

Combinations	No. of distinct units	No. of multi-copy units	Percentage of multi-copy units
Silence-Initial	22	22	100.0%
Silence-Final	49	42	85.7%
Initial-Final	616	506	82.1%
Final-Initial	1164	935	80.3%
Final-Final	1431	1242	86.8%
Final-Silence	53	53	100.0%
Overall	3335	2800	84.0%

Table 5.3 Statistics in multi-copy units using level 3 identical definition.

The availability in multiple copies would be lowered if the identical level was high. When we design the selection process, we can first decide to search for the availability of the candidate units in the most specific requirement – level 1 identical. If there is no such unit available, other identical levels, such as level 2 identical or level 3 identical, are needed to increase our chance in searching a suitable candidate.

5.1.2 Variations in acoustic parameters

In addition to the availability of multiple copies, the acoustic variations among the candidate speech segments is an important issue because they are not the exact copies of each other. If all candidates share similar properties, using any one of them would not make any differences to the result of synthesis. However, if they deviate a lot in acoustic properties, some decision criteria are needed to determine the best choice among the candidates.

We focus particularly on the match of features in pitch level, duration and intensity level between different speech segments. Although we can apply prosodic modification to adjust such features, we should try to minimize the degree of such

modification so as to avoid degradation of voice quality [2]. The acoustic features of a selected speech segment need to match with the target prosody as well as the neighboring segments. In the following sections, the variations of acoustic parameters of the sub-syllable units will be analyzed.

5.1.2.1 Pitch level

Pitch level is a critical parameter for speech synthesis of a tone language. Segments carrying different pitch contours may imply different meanings. For each syllable to be synthesized, a target pitch contour is specified. It is preferable that the sub-syllable segments selected for concatenation are at similar pitch level so that no substantial modification is needed.

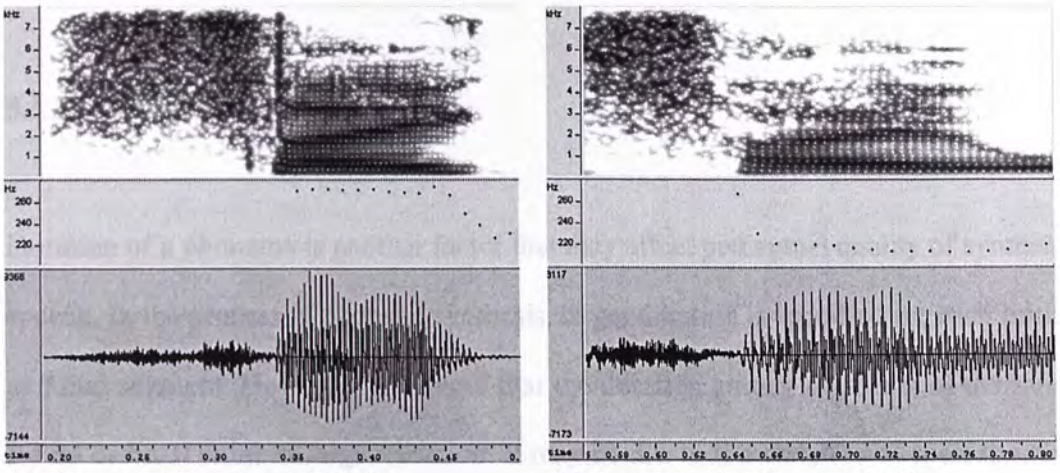
For the synthesis of a phoneme (Initial or Final), two speech segments are involved. They are typically from different utterances. A simple assumption is that matching in tones implies matching on pitch levels. However, this may not be true. We find that the speech segments have a large variation of pitch. Table 5.4 gives an example of the Final /ai/. There are 445 segments of /ai/ in the inventory, distributed over six different tones. Even within the same tone, the pitch variation can be as large as 70 Hz (Tone 1). The variation is even larger if level 2 identical or level 3 identical are used.

		No. of tokens	Pitch level (Hz)			
			Mean	Standard Deviation	Minimum	Maximum
Level 1	Tone 1	177	228.8	12.84	202.5	275.8
	Tone 2	20	157.0	6.92	146.8	173.9
	Tone 3	82	189.1	14.36	146.8	213.4
	Tone 4	93	146.6	9.07	131.1	188.3
	Tone 5	54	168.0	10.27	141.6	192.8
	Tone 6	19	165.5	6.40	152.4	181.8
Level 2	Level	371	196.2	36.51	131.1	275.8
	Rising	74	165.1	10.65	141.6	192.8
Level 3		445	191.0	35.55	131.1	275.8

Table 5.4 Distribution of pitch levels in different levels of identical for Final /ai/.

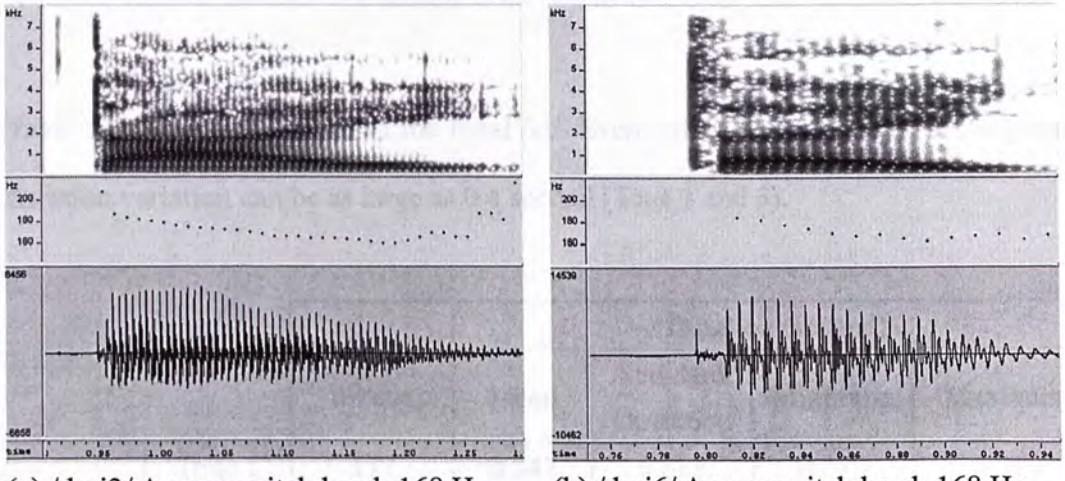
The standard deviation of pitch level for segments using level 1 identical classification (i.e. Tone 1 to 6) are roughly about 10 to 15 Hz. However, such variations increase to more than 30 Hz if other identical levels are used. In order to have a more flexible system, selection criteria for sub-syllable units in pitch levels are set to allow 30 Hz pitch difference with the target prosody.

The pitch levels of Tone 3, 4 and 6 are very close to each other. In most cases, their pitch levels fall into the range from 150 Hz to 180 Hz. This suggests that, within this group, the choice does not have to be restricted to the target tone. What should be considered are the actual pitch levels.



(a) Average pitch level: 271 Hz (b) Average pitch level: 216 Hz

Figure 5.1 Spectrogram and pitch contours for the speech segment /s-ai1/ showing same tone may have large pitch variation.



(a) /d-ai3/ Average pitch level: 168 Hz (b) /d-ai6/ Average pitch level: 168 Hz

Figure 5.2 Spectrogram and pitch contours for the speech segment /d-ai/ showing different tones may have similar pitch levels and pitch contours.

The above example shows only the case of the Final /ai/. In fact, such variations can be found in other Finals. In Appendix 1, the mean and standard deviation of pitch levels for all Finals are given. Although the concatenated speech will be subjected to prosodic modification by signal processing technique like TD-PSOLA [3] [4], the speech quality may still be highly affected. Analysis in pitch level is crucial in the selection of appropriate segments.

5.1.2.2 Duration

Duration of a phoneme is another factor that may affect perceptual quality of synthetic speech. In the process of acoustic synthesis, target duration is specified for each Initial or Final segment. However, we found that the duration among segments of the same Initial or Final exhibits large variation. It may be due to the variation of speaking rate and the recording conditions. It is preferable that the sub-syllable segments selected for concatenation have comparable duration with the target. Otherwise, unnatural speech may be resulted if the speech is too fast or too slow.

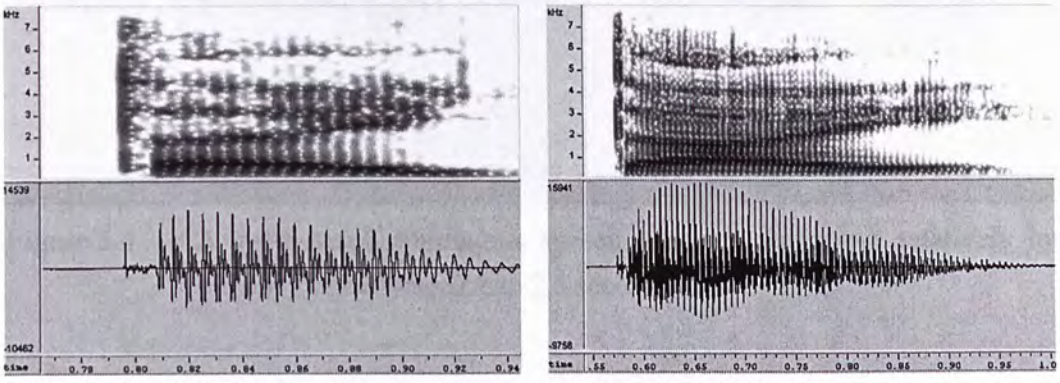
Table 5.5 gives an example of the Final /ai/. Even within the same tone, the segment duration variation can be as large as 0.4 second (Tone 1 and 3).

		No. of tokens	Duration (second)			
			Mean	Standard Deviation	Minimum	Maximum
Level 1	Tone 1	177	0.347	0.088	0.079	0.510
	Tone 2	20	0.265	0.104	0.100	0.387
	Tone 3	82	0.309	0.103	0.069	0.480
	Tone 4	93	0.291	0.066	0.109	0.426
	Tone 5	54	0.274	0.096	0.111	0.380
	Tone 6	19	0.363	0.110	0.098	0.473
Level 2	Level	371	0.325	0.091	0.069	0.510
	Rising	74	0.271	0.097	0.100	0.387
Level 3		445	0.316	0.094	0.069	0.510

Table 5.5 Distribution of duration in different levels of identical for Final /ai/.

The standard deviation at all identical levels are roughly about 0.1 second, which is about one third of the mean duration. Selection criteria for sub-syllable units in

duration will be set to allow a 30% difference in duration when compared with the target value.



(a) Duration: 0.098 second (b) Duration: 0.440 second

Figure 5.3 Spectrogram for the speech segment /d-ai6/ showing large variations in duration for the same phonetic contents.

The mean and standard deviation of segment duration of other Initials and Finals are given in Appendix 2.

5.1.2.3 Intensity level

Compared with the pitch level and duration, the intensity level of a speech segment is less directly related to speech understanding. However, abrupt change of intensity between the concatenated segments would lead to poor perceptual quality of the synthesized speech. Intensity is considered to be closely related with the perceived loudness [5]. If the intensity of a speech segment is low, we may imply that the perceived speech is weak in loudness. A natural speech utterance perceived by human normally kept in similar level of loudness. As a result, control in intensity level between speech segments is needed in synthesizing a speech utterance.

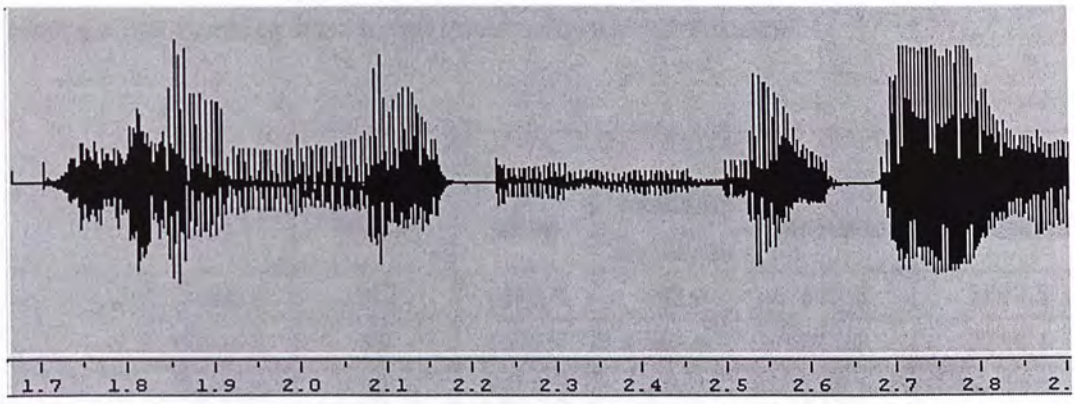


Figure 5.4 A synthesized continuous speech segment showing a relatively low intensity in between 2.2 to 2.5 seconds.

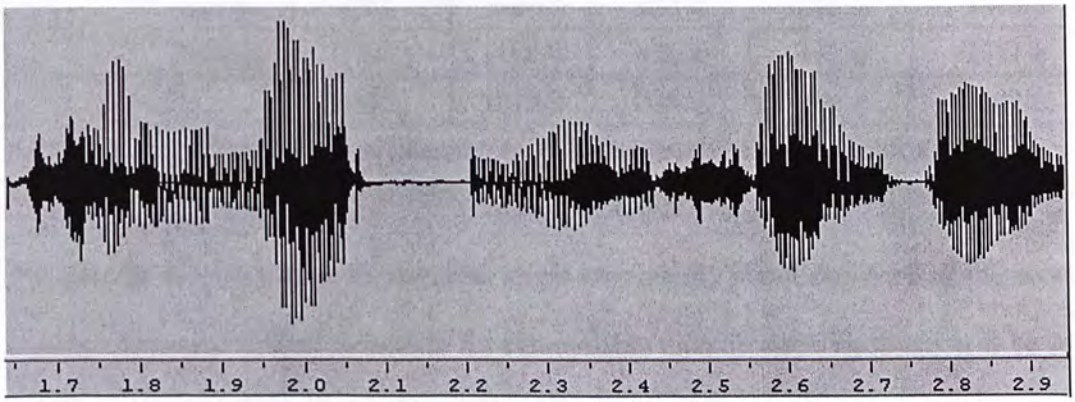


Figure 5.5 A synthesized continuous speech with improved intensity levels by replacing another speech segment with the same phonetic contents.

Unlike pitch levels and duration, no specification on intensity levels is given in our Cantonese TTS system. The sub-syllable segments are concatenated at overlapped areas. Each of the source segments will contribute part of the synthesized segment. It is likely that there exists a mismatch of intensity between the two source segments. We need to ensure a smooth intensity change and avoid abrupt change of perceived loudness. It is found that the intensity level varies greatly among the speech segment carrying the same phonemes. Here, the intensity level of a speech segment is defined as the overall average of the samples' amplitudes within the duration of the segment. Table 5.6 gives an example of the Final /ai/. Average intensity level in different

identical levels varies from several hundred to several thousand.

		No. of tokens	Intensity			
			Mean	Standard Deviation	Minimum	Maximum
Level 1	Tone 1	177	1482.6	502.9	572.6	3899.5
	Tone 2	20	1203.4	463.6	561.2	2258.1
	Tone 3	82	1513.2	474.7	680.2	3433.9
	Tone 4	93	788.8	395.4	246.6	2146.4
	Tone 5	54	1156.1	413.9	431.6	2371.8
	Tone 6	19	1204.4	463.8	670.1	2350.3
Level 2	Level	371	1301.2	557.8	246.6	3899.5
	Rising	74	1168.9	425.1	431.6	2371.8
Level 3		445	1279.2	539.9	246.6	3899.5

Table 5.6 Distribution of intensity in different levels of identical for Final /ai/.

The standard deviation at all identical levels are roughly about one third of the mean average intensity. Selection criteria for sub-syllable units in intensity levels will be set to allow a 30% difference between the segments to be concatenated. Appendix 3 shows the mean and standard deviation of average intensity levels for all Initials and Finals stored in the acoustic inventory.

5.2 Selection process: availability check on sub-syllable units

Given a list of sub-syllable units (generated by the text-processing module), the unit selection process is done sequentially, i.e. the first sub-syllable unit in the sequence will be process first and then the second one. In other words, the decision for a particular unit in the sequence would depend on that for its preceding unit. There are two stages of selection process, availability check and acoustic analysis. In availability check, we search for the possible candidates that may be suitable in

synthesizing target sub-syllable units. If more than one choice is found, acoustic analysis will start to select the most suitable segment according to their acoustic properties by some rules. Details of the analysis will be discussed in next section.

Basically, the availability check will go through at most three times for each sub-syllable units. The search start to identify segments with level 1 search, which means we use definition of level 1 identical to match the units inside the inventory with the target unit. After searching, either one of the following three results will be provided. They are "multiple copies found", "unique copy found" and "no matched copy found". Different procedures or another search using definition of level 2 identical or level 3 identical may be applied according to the results in different searches. A flow diagram of the availability check in selection process is attached below for reference.

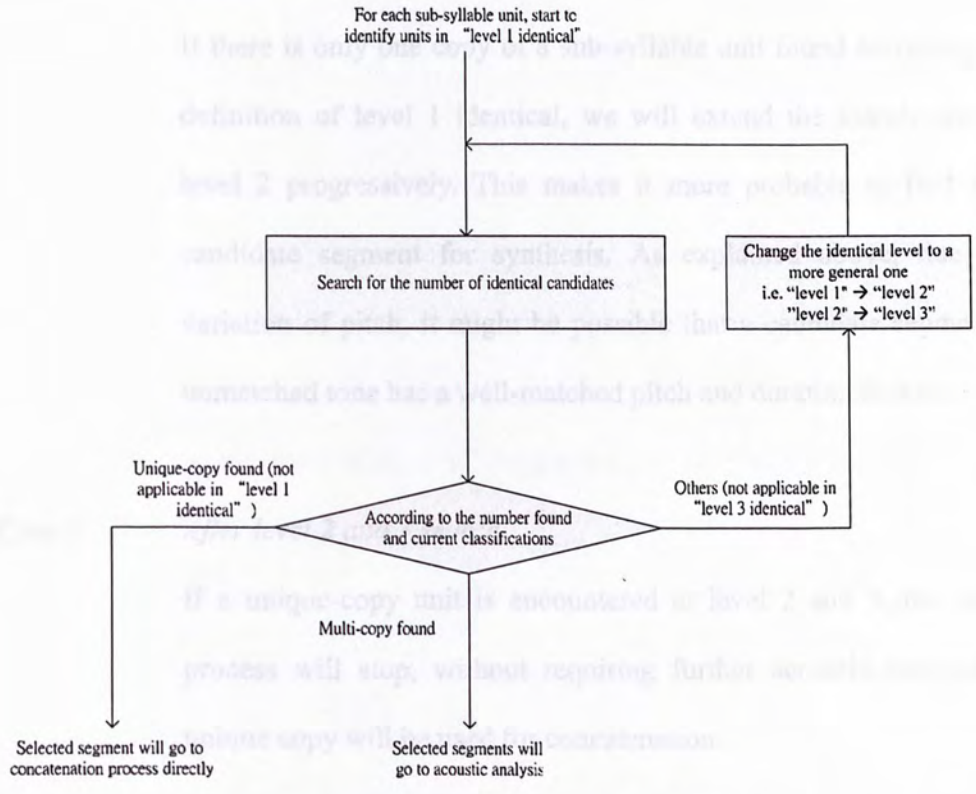


Figure 5.6 Flow diagram of the availability check in selection process.

5.2.1 Multiple copies found

The availability check will stop once there are multiple copies found for the wanted unit. All these copies are considered as the candidates for subsequent selection. These segments will proceed to the acoustic analysis stage.

5.2.2 Unique copy found

If only one copy of the speech segment is found at the current search level, two different approaches will be used as described below.

Case 1: *After level 1 search*

If there is only one copy of a sub-syllable unit found according to the definition of level 1 identical, we will extend the search process to level 2 progressively. This makes it more probable to find a good candidate segment for synthesis. As explained above, due to the variation of pitch, it might be possible that a candidate segment with unmatched tone has a well-matched pitch and duration feature.

Case 2: *After level 2 and 3 search*

If a unique-copy unit is encountered at level 2 and 3, the selection process will stop, without requiring further acoustic analysis. This unique copy will be used for concatenation.

5.2.3 No matched copy found

At level 1 and 2 search, it is possible that we cannot find any speech segment that satisfies the identical requirement. In this case, we will proceed to the next level for further search. The design of the inventory guarantees that at least one copy of the speech segment should be found from the inventory if we do not care about the tone identity.

5.2.4 Illustrative examples

Example 1: /d-ak6/

After level 1 search in availability check, there are 14 possible candidates found. They share exactly the same phonemic and tone

identities. All these candidates will pass to the acoustic analysis stage.

Example 2: /k-au1/

After level 1 search in availability check, no matched copy is found. Then we proceed to relax the tone by level 2 search. The availability check starts with definition in level 2 identical. There are 15 copies found in the inventory (including 6 copies of /k-au3/ and 9 copies of /k-au4/). All these candidates will pass to the acoustic analysis stage.

Example 3: /b-aa6/

After level 1 search, only one copy is found. To include more possible candidates for the selection process, we proceed to level 2 search and subsequently 10 copies are found (including 8 copies of /b-aa1/, 1 copy of /b-aa3/ and 1 copy of /b-aa6/). All these candidates will pass to the acoustic analysis stage.

5.3 Selection process: acoustic analysis on candidate units

Acoustic analysis is needed when there are multiple candidate segments found for a specific sub-syllable unit. These segments may carry exactly the designated tone, or a tone of the same class, or even a completely different tone. During the analysis, several acoustic parameters are considered and compared between candidate segments and prosodic specification. Parameters from target prosody are the results of statistical analysis from a large Cantonese speech database [6]. At the same time, we need to compare the acoustic parameters between adjacent segments so that they are close to each other. The general decision criteria are described as follows.

Pitch level

Since many Initials are transient, it is difficult to extract accurate pitch levels for Initials. Pitch levels of Finals will be checked in our acoustic analysis. Basically, the acoustic analysis is focused on the following two aspects:

1) Pitch difference between the candidates and the target prosody;
2) Pitch difference between the candidates and previously selected preceding segment.

It was shown earlier that there exists large variation in pitch levels even for the segments having exactly the same tone. In order to select the most suitable segment which is close to the target prosody and other speech segments, we allow at most 30 Hz of pitch level difference between them.

Duration

All Initials and Finals will undergo the duration test. All units have its own target duration specified by the prosody model. We calculate the duration difference between the candidates and the target. According to the results, we choose those candidates that have an acceptable difference in duration. We allow at most 30% of the difference between target duration and segment duration.

Intensity level

Currently, the prosody model of our TTS system does not specify the target intensity level. Thus we are mainly concerned with the intensity level between the candidates and previously selected preceding segment. The intensity level between neighboring segments should be close. As mentioned in Section 5.1.2.3, we allow at most 30% of the difference in average intensity between a candidate segment of the current unit and the preceding unit. Such difference corresponds to approximately one standard deviation from the mean value.

According to the general conditions listed above, testing parameters used in the acoustic analysis stage are summarized in the following tables according to positions and types of the units.

Position of phoneme	Unit type	Rules	
		Code	Details of the rule
pa	Initials	AI1	Percentage of difference of the intensity levels between *pa* of the current unit and *pb* of the preceding unit is smaller than 30%
		AI2	Percentage of difference of the durations between *pa* of the current unit and target duration is smaller than 30%
	Finals	AF1	Absolute difference of pitch levels between *pa* of the current unit and *pb* of the preceding unit is smaller than 30Hz
		AF2	Percentage of difference of the intensity levels between *pa* of the current unit and *pb* of the preceding unit is smaller than 30%
		AF3	Absolute difference of pitch levels between *pa* of the current unit and target pitch is smaller than 30Hz
		AF4	Percentage of difference of durations between *pa* of the current unit and target duration is smaller than 30%
	Silence	AS1	Intensity levels within this segment is low (generally under 25)
pb	Initials	BI1	Percentage of difference of durations between *pb* of the current unit and target duration is smaller than 30%
	Finals	BF1	Absolute difference of pitch levels between *pb* of the current unit and target pitch is smaller than 30Hz
		BF2	Percentage of difference of durations between *pb* of the current unit and target duration is smaller than 30%
	Silence	(No parameters to be tested)	

Table 5.7 Testing parameters to be used according to their positions and units.

There are six different types of sub-syllable units (see Table 3.1). For each type, different rules listed in the above table will be considered together. Generally, it is expected that the selected speech segment can fulfill all requirements in pitch, duration and intensity. However, if none of the candidates can satisfy all conditions,

the conditions on pitch would be given a higher priority. The following table is a summary on the priorities used in different conditions and units during the analysis.

Combinations (*pa-pb*)	Priorities	Conditions required to fulfill
Silence–Initial	(1)	Fulfill AS1 and BI1.
	(2)	If no unit can fulfill (1), then unit with smallest difference in condition BI1 is selected.
Silence–Final	(1)	Fulfill conditions AS1, BF1 and BF2.
	(2)	If no unit can fulfill (1), then unit with smallest difference in condition BF1 is selected.
Initial–Final	(1)	Fulfill conditions AI1, AI2, BF1 and BF2.
	(2)	If no unit can fulfill (1), then check for units to fulfill [AI2, BF1 and BF2] or [AI1 and BF1]
	(3)	If still no unit can fulfill (2), then unit with smallest difference in condition BF1 is selected.
Final–Initial	(1)	Fulfill conditions AF1, AF2, AF3, AF4 and BI1.
	(2)	If no unit can fulfill (1), then check for units to fulfill [AF1, AF2 and AF3] or [AF1, AF3, AF4 and BI1]
	(3)	If still no unit can fulfill (2), then check for units to fulfill [AF1 and AF3]
	(4)	If still no unit can fulfill (3), then unit with smallest difference in condition AF3 is selected.
Final – Final	(1)	Fulfill conditions AF1, AF2, AF3, AF4, BF1 and BF2.
	(2)	If no unit can fulfill (1), then check for units to fulfill [AF1, AF2, AF3 and BF1] or [AF1, AF3, AF4, BF1 and BF2]
	(3)	If still no unit can fulfill (2), then check for units to fulfill [AF1, AF3 and BF1]
	(4)	If still no unit can fulfill (3), then unit with smallest difference in sum of AF3 and BF1 is selected.
Final–Silence	(1)	Fulfill conditions AF1, AF2, AF3 and AF4.
	(2)	If no unit can fulfill (1), then check for units to fulfill [AF1, AF2 and AF3] or [AF1, AF3 and AF4]
	(3)	If still no unit can fulfill (2), then check for units to fulfill [AF1 and AF3]
	(4)	If still no unit can fulfill (3), then unit with smallest difference in condition AF3 is selected.

Table 5.8 Table shown the priorities used in different conditions and units during the analysis.

On each priority level, if only one candidate passed the required conditions, then it is no doubt to select the passed segment for further processing. However, if there exist more than one segments fulfill the required conditions, then the segment with the smallest difference of pitch levels between candidate segment and target pitch is

selected.

Selected examples

Example 1: /s-aa1/

After the availability check, 4 candidates are found after level 1 search. Their acoustic properties are examined as listed in the following table.

(In the following table, "+" means sample value is larger than the target while "-" means sample value is smaller than the target. Fail cases are marked with gray boxes.)

(Code)	AI1		AI2		BF1		BF2	
Target value / preceding segment's value	Average intensity: 758.5		Duration: 0.123		Average pitch: 227.8 Hz		Duration: 0.210 (s)	
Candidate	Average intensity	Difference (in percentage)	Duration (s)	Difference (in percentage)	Average pitch (Hz)	Difference (Hz)	Duration (s)	Difference (in percentage)
Candidate 1	668.5	-12%	0.094	-24%	216.2	-11.6	0.200	-5%
Candidate 2	605.0	-20%	0.099	-20%	207.8	-20.0	0.404	+92%
Candidate 3	151.8	-80%	0.137	+11%	250.1	+22.3	0.133	-37%
Candidate 4	272.5	-64%	0.112	-9%	216.2	-11.6	0.133	-37%

Table 5.9 Summary in the testing conditions with the acoustic parameters used in example 1.

Candidate 1 fulfilled all the conditions required in selecting this target sub-syllable unit. So it is selected for the synthesis purpose.

Example 2: /aa1-t/

After the availability check, 3 candidates are found after level 1 search. Their acoustic properties are examined as listed in the following table.

(In the following table, "+" means sample value is larger than the target while "-" means sample value is smaller than the target. Fail cases are marked with gray boxes.)

(Code)	AF1		AF2		AF3		AF4		BI1	
Target value / preceding segment's value	Previous average pitch: 216.2 Hz		Previous average intensity: 1752.6		Target average pitch: 227.8 Hz		Target duration: 0.210 (s)		Target duration: 0.078 (s)	
Candidate	Average pitch (Hz)	Difference (Hz)	Average intensity	Difference (in percentage)	Average pitch (Hz)	Difference (Hz)	Duration (s)	Difference (in percentage)	Duration (s)	Difference (in percentage)
Candidate1	242.4	+26.2	1774.7	+1%	242.4	+14.6	0.198	-6%	0.100	+28%
Candidate2	250.1	+33.9	1750.1	-0%	250.1	+22.3	0.133	-33%	0.094	+21%
Candidate3	219.2	+3.0	1217.3	-31%	219.2	-8.6	0.133	-33%	0.085	+9%

Table 5.10 Summary in the testing conditions with the acoustic parameters used in example 2.

Statistics shown that candidate 3 contains the smallest difference in pitch levels both compared with target pitch and preceding segment. However, its intensity levels and durations are in large variations. On the other hand, candidate 1 can fulfill all the requirements that the differences are in acceptable range. In this example, candidate 1 is selected for the synthesis purpose.

References:

[1] K.M. Law, "Cantonese Text-to-Speech Synthesis Using Sub-syllable Units", M. Phil. Thesis, the Chinese University of Hong Kong, 2001

[2] S. Narayanan & A. Alwan, "Text-to-Speech Synthesis: New Paradigms and Advances", New Jersey: Pearson Education, Inc, 2005

[3] E. Moulines & J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech", Speech Communications Vol. 16, pp. 175-205.

[4] E. Moulines & F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communications Vol. 9, pp. 453-467.

[5] Y. J. Li, "Prosody Analysis and Modeling for Cantonese Text-to-Speech", M. Phil. Thesis, the Chinese University of Hong Kong, 2003

[6] Tan Lee, Helen Meng, W. Lau, W. K. Lo & P. C. Ching, "Micro-prosodic control in Cantonese Text-to-Speech synthesis", in Proc. Eurospeech-99, Vol. 4, pp. 1855-1858, Budapest.

Chapter 6

Performance Evaluation

We have presented the complete set of strategies of unit selection and waveform concatenation for sub-syllable based Cantonese TTS. In this chapter, the proposed methods will be evaluated from different perspectives.

The performance evaluation was divided into two parts. The first part was an objective test. In this test, we attempted to assess the acoustic properties of the selected sub-syllable units. The second part was a subjective perceptual test. A number of human subjects participated into the test in which they listened to and graded the synthesized speech. The same set of test materials were used in both tests.

We compared the performance of the system using the proposed methods with the baseline system as described in [1]. In the baseline system, there was no unit selection process. All sub-syllable units were specified by a prescribed list without considering the actual acoustic properties of the speech segments. The waveform concatenation strategies were very simple, as mentioned in Chapter 3.3.

In the whole evaluation process, other modules of the system, namely text processing and prosodic modification, were kept as the same.

6.1 General information

6.1.1 Objective test

In this test, we measured the acoustic difference between the selected sub-syllable segments and the target prosody. Two acoustic parameters, namely pitch level and segment duration, were concerned in the test. The target pitch and duration were the results of statistical analysis from a large database [2]. They are given as in Appendix 6 and 7.

6.1.2 Subjective test

The objective test can provide us with quantitative assessment on the acoustic match or mismatch. However, improvement in measurable data may not directly reflect the change of perceptual quality. To have a complete overview of the Cantonese TTS system, we need the response and feedback from human users. It is an important factor because user satisfaction is the most important consideration in the design of human-computer interaction process.

In the subjective perceptual test, sample speeches are synthesized by the two systems. Human subjects are asked to grade the samples based on their perceptual feeling. There are a large number of subjects so that their assessment results can be analyzed statistically to give an overall evaluation of the system.

6.1.3 Test materials

The test materials contained words and paragraphs. All the words, sentences and paragraphs were in the form of written Chinese and extracted from several local newspapers on various topics.

Words

There were altogether 50 words to be tested. Each word contains two to four Chinese characters (syllables). Different types of Initials and Finals are included so that different concatenation strategies and unit selection criteria can be evaluated. The coverage of the word list is described as in Table 6.1. The words are given as in Appendix 4.

Total number of words	50
Number of words in 2 syllables	35
Number of words in 3 syllables	10
Number of words in 4 syllables	5
Total number of Initial segments	119
Plosive and Affricate	45
Fricative	26
Voiced Initial (Nasal and others)	48
Total number of Final segments	120
Long Vowel and Nasal	20
Diphthong	38
Vowel with Nasal Coda	44
Vowel with Stop Coda	18

Table 6.1 Summary of the coverage of the word list.

Paragraphs

There were 25 paragraphs selected. Each contains one to three sentences and each sentence has 4 to 37 Chinese characters. The coverage of these paragraphs is given as in Table 6.2. The complete contents are given in Appendix 5.

Number of paragraphs	25
Number of sentences	40
Number of Chinese Characters included	597
Average number of Characters per sentence	15

Table 6.2 Summary of the coverage of the paragraph list.

6.2 Details of the objective test

The first part of the test basically focuses on the performance in unit selection process. We believe that the sub-syllable units selected via the proposed selection process will be better matched with the target prosody when compared with the baseline system.

6.2.1 Testing method

All the 120 syllables in the test words and the 597 syllables in the test paragraphs were used in this objective test. We measured the pitch level and the segment duration of their Final segments. We focus on the Final segments only because they are the core of the Cantonese syllables. There are altogether 240 sub-syllable segments in the test words and 1194 sub-syllable segments in the test paragraphs that contain Finals.

The test words and paragraphs were synthesized by the two systems. During the

synthesis process, we recorded the sub-syllable segments selected for synthesis and find out the pitch and segment duration for each Final segment. These values were compared with the pitch and duration as given by the target prosody, and their discrepancies were analyzed.

6.2.2 Results

We present the results of analysis in four parts. They are: 1) pitch difference for the test words; 2) pitch difference for the test paragraphs; 3) duration difference for the test words; and 4) duration difference for the test paragraphs.

Pitch difference (words)

The average pitch differences between the selected segments and the target prosody in the synthesis of test words are summarized as in Table 6.3. For example, there are 52 segments selected with the absolute pitch difference of 20 Hz or more. Moreover, there are 82 segments selected with the absolute pitch difference of 15 Hz or more, including the 52 segments mentioned before. The overall mean of the absolute pitch difference drops from about 16 Hz (baseline system) to about 10 Hz (proposed system). It shows that the proposed unit selection method indeed outperform the baseline system, by providing a better matched in pitch values. The percentage of speech segments having absolute pitch difference of 10 Hz or more drops from 58% to about 37%. For the cases in absolute pitch difference of 20 Hz or more, it drops from 21% to about 10%. Nevertheless, the proposed methods cannot solve all problems because some of the sub-syllable units have only one copy in the acoustic inventory.

Absolute pitch difference	Baseline system		Proposed system	
	Counts	Percentage (%)	Counts	Percentage (%)
20Hz or more	52	21.7	24	10.0
15Hz or more	82	34.2	49	20.4
10Hz or more	140	58.3	89	37.1
Mean	16.45 Hz		9.46 Hz	
SD	16.62 Hz		9.97 Hz	
Maximum difference	79.00 Hz		55.78 Hz	

Table 6.3 Summary of the absolute pitch difference using test words.

Pitch difference (paragraphs)

The average pitch differences between the selected segments and the target prosody in the synthesis of test paragraphs are summarized as in Table 6.4. The pitch levels of the selected speech segments are better matched to the target prosody, similar to the results by test words. The overall mean of the absolute pitch difference between the target pitch and the segment pitch drops from 20 Hz to 10 Hz. The percentage of speech segments selected with large discrepancies is significantly reduced. For example, the percentage of speech segments having absolute difference of 10 Hz or more drops from 60% to 36%. For the case in absolute difference of 20 Hz or more, it drops from 32% to 16%. It shows that selection process can help us in choosing better matched segments in pitch levels.

Absolute pitch difference	Baseline system		Proposed system	
	Counts	Percentage (%)	Counts	Percentage (%)
20Hz or more	388	32.5	198	16.6
15Hz or more	521	43.6	394	24.6
10Hz or more	720	60.3	430	36.1
Mean	20.22 Hz		10.71 Hz	
SD	21.01 Hz		12.20 Hz	
Maximum difference	112.65 Hz		107.83 Hz	

Table 6.4 Summary of the absolute pitch difference using test paragraphs.

To summarize, the proposed unit selection method can choose speech segments with a better matched in pitch levels.

Segment duration (words)

The duration difference between the selected segments and the target prosody for the test words are summarized as in Table 6.5. The overall mean of duration difference between the target values and the segments has small improvement (from 0.097 second to 0.080 second) over the baseline system. However, the cases of large discrepancies are greatly reduced. For example, the percentage of speech segments having absolute difference of 0.15 second or more drops from 26% to 15%. Moreover, the percentage of speech segments having absolute difference of 0.2 second or more drops from 11% to 2%.

Absolute duration difference	Baseline system		Proposed system	
	Counts	Percentage (%)	Counts	Percentage (%)
0.20s or more	26	10.8	5	2.1
0.15s or more	64	26.6	37	15.5
0.10s or more	95	39.6	75	31.3
Mean	0.097 second		0.080 second	
SD	0.076 second		0.054 second	
Maximum difference	0.312 second		0.246 second	

Table 6.5 Summary of the absolute duration difference using test words

Segment duration (paragraphs)

Table 6.6 shows the results of segment duration on the test paragraphs. Similar to the results on words, the overall mean of duration difference has only a small improvement (from 0.078 second to 0.059 second), while the severely mismatched cases are improved. For example, the percentage drops from 26% to 18% for the cases of difference of more than 0.1 seconds.

Absolute duration difference	Baseline system		Proposed system	
	Counts	Percentage (%)	Counts	Percentage (%)
0.20s or more	100	8.4	39	3.3
0.15s or more	181	15.1	89	7.4
0.10s or more	314	26.3	221	18.5
Mean	0.078 second		0.059 second	
SD	0.085 second		0.053 second	
Maximum difference	0.312 second		0.277 second	

Table 6.6 Summary of the absolute duration difference using test paragraphs.

6.2.3 Analysis

In general, both tests show that the proposed methods of unit selection provide a

better match in pitch level and duration between the selected speech segments and the target prosody. However, acoustic properties of speech segments may not fulfill both conditions at the same time. As our selection process gives a higher priority to the pitch parameter, there may have a chance to select a segment which its segment duration is far away from the target while its pitch level is closed to the target. Result shows that both the mean difference and the standard deviation are improved by using the proposed system. It shows that the proposed unit selection method can choose better candidates from the acoustical point of view.

Nevertheless, in some cases, there are significant mismatches between the selected segments and the target prosody, in terms of both pitch and duration. There are two reasons for this. Firstly, some of the sub-syllable unit combinations only have one copy in the acoustic inventory. We do not have any choice for these units. Secondly, there exists a significant difference between the overall pitch level of the target prosody and the speech segments stored in the inventory, as shown in Table 6.7. The average pitch varies from about 10 Hz (Tone 1, 3 and 6) to more than 20 Hz (Tone 4). Such difference is due to that the speech used for prosody modeling comes from a different speaker.

Tones	Average pitch level in target prosody	Average pitch level of the segments in acoustic inventory
Tone 1	242.8 Hz	233.9 Hz
Tone 2	186.0 Hz	171.4 Hz
Tone 3	199.8 Hz	190.8 Hz
Tone 4	169.4 Hz	149.0 Hz
Tone 5	183.8 Hz	167.7 Hz
Tone 6	188.1 Hz	176.1 Hz

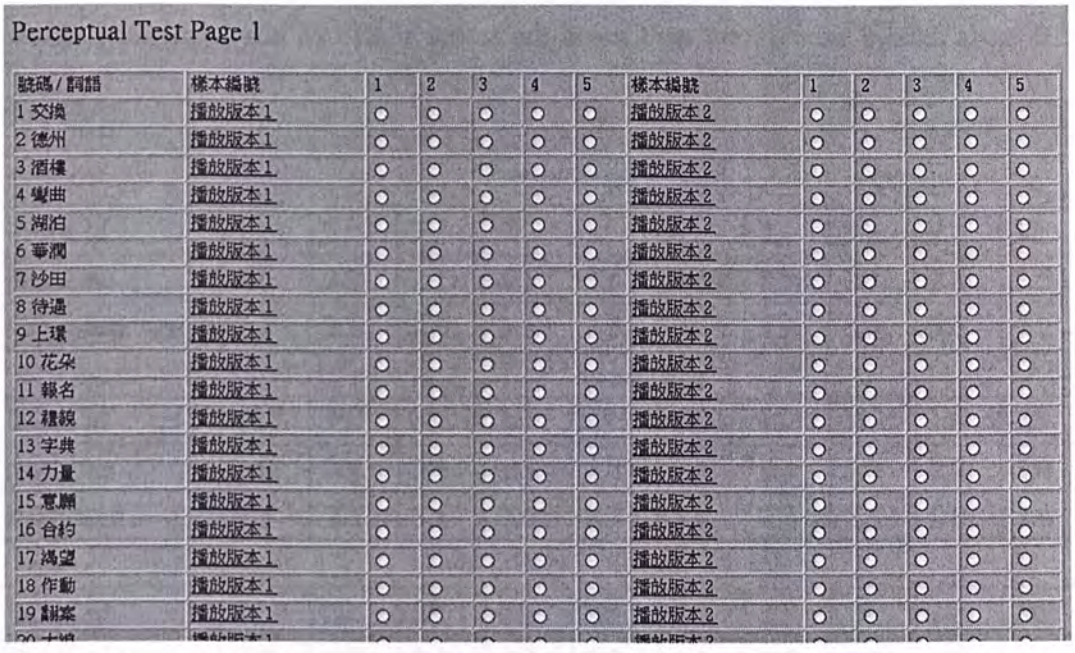
Table 6.7 Average pitch levels in both target prosody and speech segments in acoustic inventory.

6.3 Details of the subjective test

Although objective measurements show that our proposed system can help us in selecting a better-matched unit, it does not imply that it results in a better perceptual quality. Moreover, the objective test mainly evaluates the performance in unit selection only. Subjective listening tests are therefore carried out in order to collect different users' opinions on the performance of the systems.

6.3.1 Testing method

A total of 48 subjects (26 males and 22 females) were invited to do the listening test. They are all native Cantonese speakers studying in universities and post-secondary schools. We asked each subject to listen the synthesized words and paragraphs one by one. For the same word or paragraph, they heard the synthesized speech from both systems. The sequence in presenting them is randomized so that they did not know which version they were listening to. The subject was asked to grade the two utterances by 5-mark scale opinion score, in which 1 means the poorest and 5 means the best. Their scores are based on their feelings in intelligibility and naturalness of the synthesized speech. Figure 6.1 shows the computer interface for the listening test.


Perceptual Test Page 1

號碼 / 詞語	樣本編號	1	2	3	4	5	樣本編號	1	2	3	4	5
1 交換	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
2 德州	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
3 酒櫃	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
4 彎曲	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
5 湖泊	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
6 華爾	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
7 沙田	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
8 待遇	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
9 上環	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
10 花朵	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
11 報名	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
12 禮貌	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
13 字典	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
14 力量	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
15 意願	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
16 合約	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
17 渴望	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
18 作動	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
19 翻案	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○
20 大堤	播放版本 1	○	○	○	○	○	播放版本 2	○	○	○	○	○

Figure 6.1 A testing interface for the subjects to hear the speeches and give marks.

6.3.2 Results

The following table gives the mean opinion score (MOS) given by the subjects. It shows that the proposed system generally outperforms the baseline system with an absolute improvement of 0.49 for words and 0.38 for paragraphs.

System / List	Baseline system		Proposed system	
	Mean	SD	Mean	SD
Word list	2.74	0.65	3.23	0.50
Paragraph list	2.50	0.36	2.88	0.41

Table 6.8 Overall performances of the two TTS systems.

Table 6.9 shows the MOS attained for each individual word. Higher scores were given to the words synthesized by the proposed system. Some samples have the absolute improvement of 1.5 points or more, e.g. words with id 16, 33 and 49. However, we also found that the proposed system performs poor for some of the words, such as

words with id 18 and 39. Their scores are lower than the baseline system about 0.5 points.

Word id	1	2	3	4	5	6	7	8	9	10
Baseline	3.0	3.2	2.1	4.1	3.4	2.6	3.5	2.5	2.8	2.5
Proposed	3.1	3.1	2.0	3.5	3.8	3.0	3.6	3.3	3.6	3.2

Word id	11	12	13	14	15	16	17	18	19	20
Baseline	2.0	2.5	3.1	1.9	2.4	1.8	2.4	3.3	2.8	3.4
Proposed	2.5	2.5	3.6	2.9	3.3	3.8	3.5	2.6	2.9	3.6

Word id	21	22	23	24	25	26	27	28	29	30
Baseline	3.5	3.2	3.0	3.2	1.8	1.1	3.5	1.9	2.0	3.1
Proposed	3.4	3.6	4.1	3.9	3.4	3.6	3.4	2.5	3.7	4.2

Word id	31	32	33	34	35	36	37	38	39	40
Baseline	3.2	2.1	1.7	2.0	3.2	3.2	2.9	3.2	3.6	3.0
Proposed	3.4	2.5	4.0	3.0	3.7	3.3	3.5	3.0	3.0	2.9

Word id	41	42	43	44	45	46	47	48	49	50
Baseline	2.0	3.3	3.6	2.8	2.6	2.1	3.4	2.7	1.9	3.0
Proposed	2.3	3.5	3.5	2.8	2.7	2.4	2.9	2.9	3.6	3.6

Table 6.9 Detailed performances on each word in word list.

Table 6.10 shows the MOS given to individual test paragraphs. Higher scores were given to the paragraphs synthesized by the proposed system. Some utterances have an absolute improvement of 1 point or more.

Paragraph id	1	2	3	4	5	6	7	8	9	10
Baseline	1.6	2.1	2.8	2.8	2.9	2.2	2.3	2.1	2.6	2.6
Proposed	2.9	2.4	2.7	3.4	2.8	2.8	2.8	3.9	2.9	3.8

Paragraph id	11	12	13	14	15	16	17	18	19	20
Baseline	1.9	2.6	2.4	2.5	2.8	2.5	2.7	2.4	2.8	2.4
Proposed	2.2	2.9	3.0	2.5	2.4	2.7	2.7	2.7	2.8	2.9

Paragraph id	21	22	23	24	25
Baseline	3.2	2.8	2.1	2.4	2.9
Proposed	2.7	2.9	2.7	3.6	2.9

Table 6.10 Detailed performances on each paragraph in paragraph list.

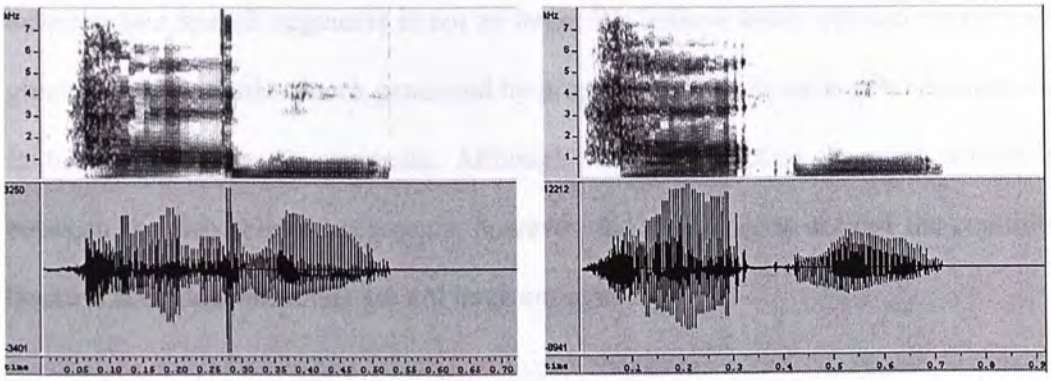
6.3.3 Analysis

In this section, we select a few example of the synthesized speech and look into details of their test results.

Case 1: Word id 33

As shown in Table 6.9, this sample gives an improvement of 2.3 points. The waveforms and the corresponding spectrograms are shown in Figure 6.2. In the baseline system, the concatenation point for the Initial /w/ is wrongly determined, causing a spectral discontinuity in the output speech. The problem is solved in the proposed system. Moreover, a different segment is chosen for the unit /aak3-w/. It shows that the pitch difference between the target pitch and the speech segments is smaller than the segments used in baseline system. The duration of the newly selected segments is not very good when compared with the units used in the baseline system. However, the overall quality of synthetic speech is much improved because of

improved concatenation technique and better selected unit.



(a) Synthesized by baseline system (b) Synthesized by proposed system

Figure 6.2 Speech waveforms and spectrograms for the utterance /haak3-wu6/.

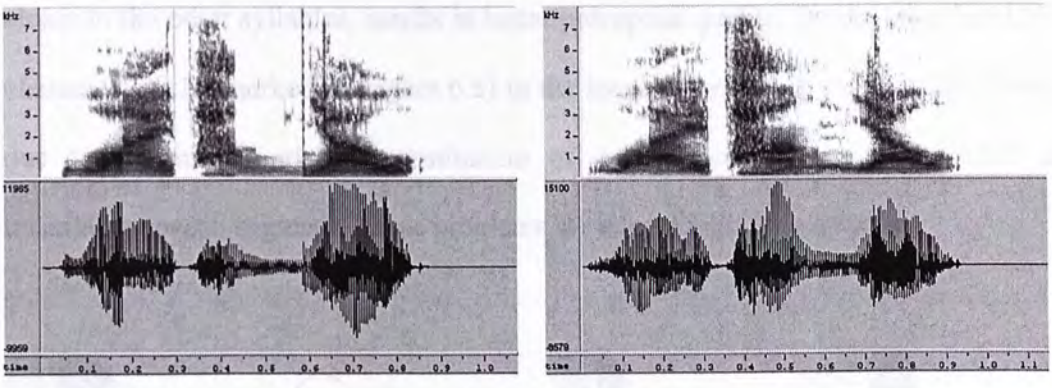
Acoustic contents of /aak3/ in sub-syllable segments	Average pitch difference	Segment duration difference
/h-aak3/ used in baseline system	14.6 Hz	0.008 second
/h-aak3/ used in proposed system	14.6 Hz	0.008 second
/aak3-w/ used in baseline system	17.7 Hz	0.047 second
/aak3-w/ used in proposed system	3.0 Hz	0.096 second

Table 6.11 Acoustic variations of Final segment /aak3/ in different speech segments

Case 2: Word id 39

This sample shows that there is a drop of 0.6 point according to the subjects' opinion. The waveforms and the corresponding spectrograms are shown in Figure 6.3. The spectrograms show that some of the spectral mismatch points are removed in the proposed system. However, the acoustic properties of the segments selected for concatenation have large variations. In sub-syllable unit /aai6-p/, the proposed system selects the segment with a higher pitch difference but with a better match in segment duration; while for the sub-syllable unit /w-aai6/, the proposed system selects the

segment with better match in pitch levels. As a result, the pitch difference between two speech segments is more than 10 Hz. In the baseline system, the pitch difference between two speech segments is not so large. We believe lower opinion scores were given to the synthetic speech generated by proposed system because of such mismatch in between the speech segments. Although we have checked the pitch difference between the two selected segments, however, the system does not fail the condition because the pitch variations are not large enough.



(a) Synthesized by baseline system (b) Synthesized by proposed system

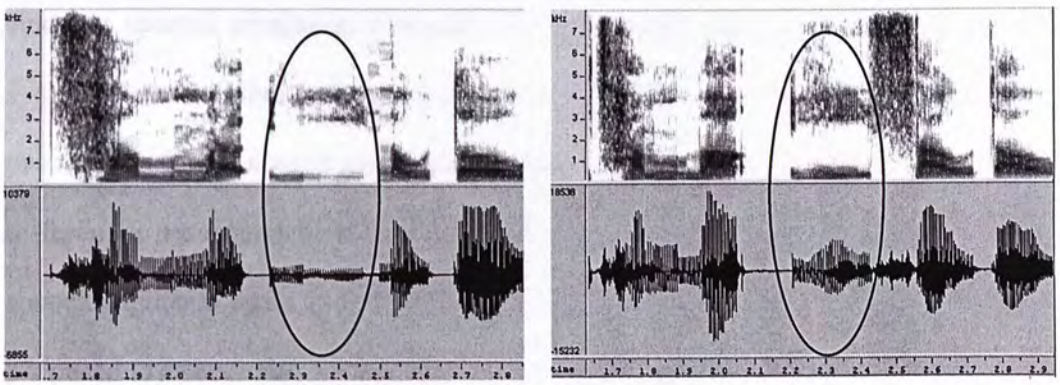
Figure 6.3 Speech waveforms and spectrograms for the utterance /waai6-pang4-jau5/.

Acoustic contents of /aai6/ in sub-syllable segments	Average pitch difference	Segment duration difference in percentage
/w-aai6/ used in baseline system	14.7 Hz	28%
/w-aai6/ used in proposed system	4.7 Hz	39%
/aai6-p/ used in baseline system	11.5 Hz	39%
/aai6-p/ used in proposed system	18.9 Hz	28%

Table 6.12 Acoustic variations of Final segment /aai6/ in different speech segments

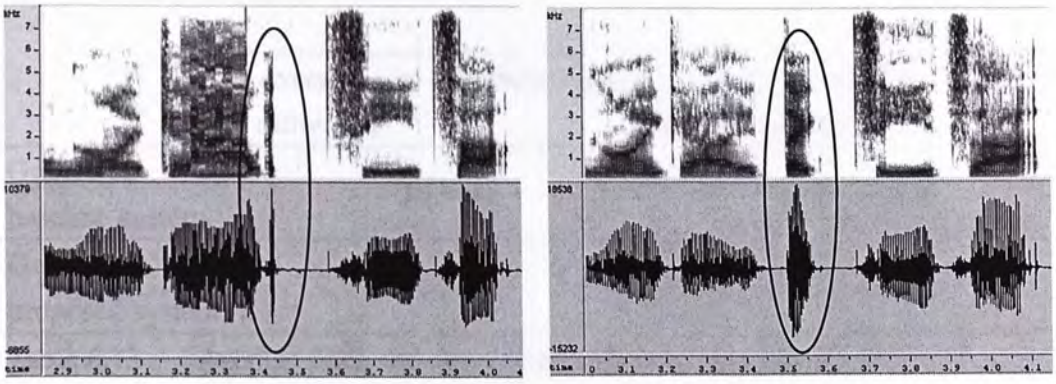
Case 3: Paragraph id 8

This sample shows that there is an improvement of 0.6 point. Parts of the waveforms and the corresponding spectrograms are shown as in Figures 6.4 and 6.5. We believe that higher in opinion scores is due to the improvement on acoustic properties in average intensity and the concatenation strategy. In the baseline system, the intensity of the syllable /ji4/ (marked in Figure 6.4) is very weak. But in the proposed system, other segments are selected in synthesizing this syllable so that the loudness is much closer to the other syllables, results in better perceptual quality. On the other hand, the character /dik1/ (marked in Figure 6.5) in the baseline system is very unclear. It may due to the inappropriate determination of concatenation points or selected an unsuitable speech segment. These problems are solved in proposed system.



(a) Synthesized by baseline system (b) Synthesized by proposed system

Figure 6.4 Speech waveforms and spectrograms for part of the utterance /cam4-mak6-ji4-sau6-dou3/.



(a) Synthesized by baseline system (b) Synthesized by proposed system

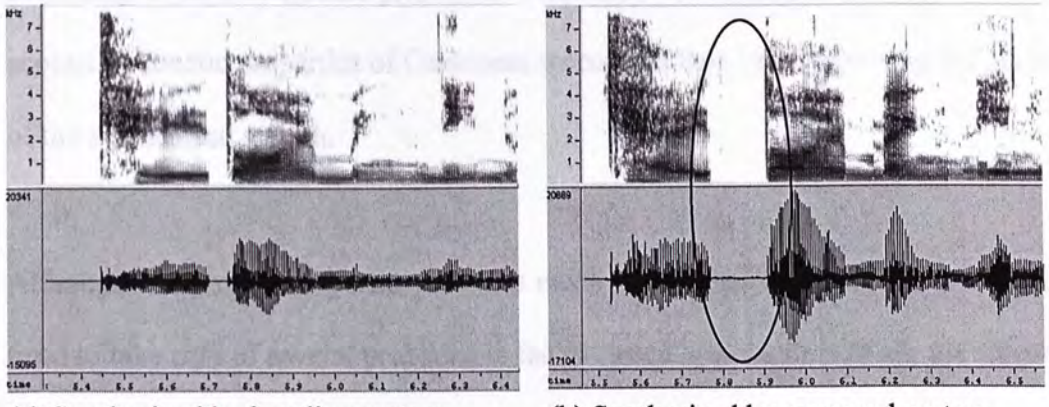
Figure 6.5 Speech waveforms and spectrograms for part of the utterance /ngoi6-gaai3-dik1-zi2-zaak6/.

Case 4: Paragraph id 21

This sample shows that there is an absolute drop in 0.5 points. Parts of the waveforms and the corresponding spectrograms are shown as in Figures 6.6 and 6.7. We tried to find out spectral mismatch points in the synthesized speech generated by proposed system. But we failed to locate such points. After that, we tried to check its acoustic properties of both speech segments. In fact, the average pitch and segment duration differences are improved as shown in table 6.13. Finally, we tried to listen with both speech segments again to find out the possibilities in giving a relatively low score in proposed system. We found that several improper pauses are appeared in the synthesized speech by the proposed system (marked in Figures 6.6 and 6.7). In continuous speech, short pauses are required to make listeners easier to interpret the contents of a long speech and identify different phrases. However, the pauses we found in the speech generated by the proposed system are exceptionally long inside a word phrase. Such pauses may obstruct the understanding of a speech utterance causing unnatural in synthetic speech. We believe that the proposed system selected inappropriate speech segments which contain phrase boundaries.

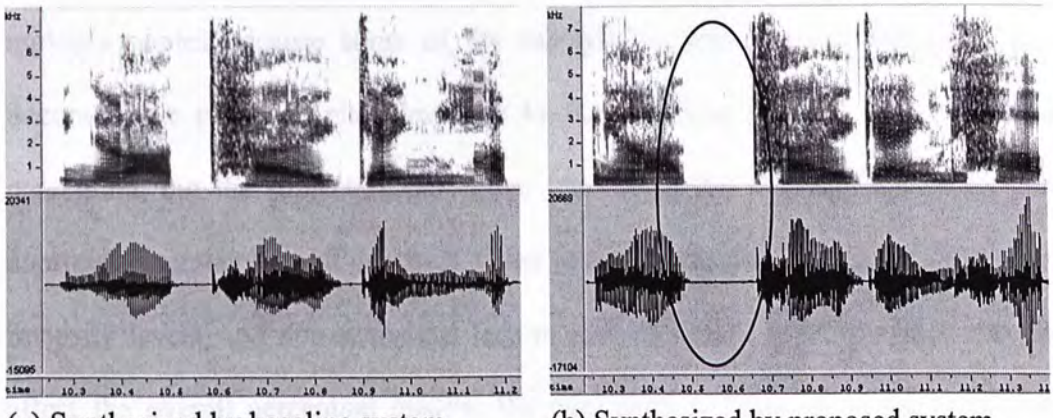
	Average in absolute pitch difference	Average in absolute segment duration difference
Generated by baseline system	16.3 Hz	0.072 second
Generated by proposed system	11.8 Hz	0.058 second

Table 6.13 Average pitch and duration difference in paragraph id 21.



(a) Synthesized by baseline system (b) Synthesized by proposed system

Figure 6.6 Speech waveforms and spectrograms for part of the utterance /kei4-doi6-ming4-nin4/.



(a) Synthesized by baseline system (b) Synthesized by proposed system

Figure 6.7 Speech waveforms and spectrograms for part of the utterance /ngaa3-taai3-ging1-hap6/.

6.4 Summary

Concatenation strategies and unit selection process proposed in the previous chapters have been proven to be effective in improving the overall performance of the TTS system. The proposed process of unit selection attempts to fully utilize the available sub-syllable units stored in the acoustic inventory. The best candidate is determined according to their acoustical properties. Proposed concatenation strategies show that acoustic-phonetic properties of Cantonese speech can help us in improving the quality of the synthesized speech.

Although the proposed system performs much better than the baseline, but we still need to take care of several problems in the proposed system. First of all, the acoustic inventory we used has a limited size. Some sub-syllable units have one choice only. We need to increase the inventory size if we want the selection process to be available for all units' combination. However, it may not be necessary for all units having multiple copies because some of the sub-syllable units are not frequently used. Secondly, we proposed pitch levels to be the dominant factor in affecting human perception, but the subjective perceptual test shows that this may not be the most appropriate strategy for all the cases. Other acoustical factors, such as the duration and intensity levels, and non-acoustical factors such as pauses in an utterance may also affect the overall perceptual quality. We may need to know more the relationship between different parameters in affecting perceptual quality of an utterance.

References:

[1] K.M. Law, "Cantonese Text-to-Speech Synthesis Using Sub-syllable Units", M. Phil. Thesis, the Chinese University of Hong Kong, 2001

[2] Tan Lee, Helen Meng, W. Lau, W. K. Lo & P. C. Ching, "Micro-prosodic control in Cantonese Text-to-Speech synthesis", in Proc. Eurospeech-99, Vol. 4, pp. 1855-1858, Budapest.

Chapter 7
Conclusions and Future Works

7.1 Conclusions

Cantonese TTS has been studied for some years by our research group. The quality of synthetic speech is not in high standard, in terms of intelligibility and naturalness. Sub-syllable based system was proposed to improve the speech quality by retaining the co-articulation between syllables. However, for corpus-based concatenative approach, the quality of the synthesized speech depends greatly on the speech segments selected and the way of concatenating them. In this thesis, we aim to develop appropriate methods of unit selection and waveform concatenation for Cantonese TTS. Our investigation is focused on the acoustic inventory of sub-syllable units developed by Law [1]. Design of a new acoustic inventory is not the concern of this thesis. Nevertheless, we believe that the methods will be generally applicable to any corpus-based Cantonese speech synthesis systems.

Concatenation strategies are one of the key elements in a corpus-based system. Since the speech segments to be concatenated are from different linguistic and acoustic contexts, we need to find a way to minimize the artifacts generated in the vicinity of the concatenation points. Based on the acoustic-phonetic properties, the Cantonese phonemes are classified into different groups. Initials are classified according to their manner of articulation. Finals are classified according to their phonetic structure. Speech units in the same group share similar properties that are important in concatenation. Concatenation strategies are proposed for different groups. Generally the

process of concatenation is divided into two steps: 1) determination of the concatenation point in each of the speech segments; 2) concatenation of the waveforms in time domain and smoothing.

Unit selection was not a consideration in the existing version of the TTS system developed by Law [1]. It simply uses a selection list in which each sub-syllable unit is mapped to only one specific speech segment. This segment is used regardless of its context. Our statistical analysis shows that there are quite a large number of synthesis units that are available in the acoustic inventory, but never used for the synthesis. To fully utilize these units, a unit selection process is proposed. The main idea is to allow the largest number of candidates stored inside the inventory and select the best one from the acoustical point of view. We define different levels of similarity between the candidate segments. At the lowest level, the segments are required to have the same phonemic elements and the tones are not considered. At the highest level, we require that the tones have to be identical. At different levels, we may find multiple candidates for a desired unit. The most appropriate candidate is selected based on the match of acoustic properties with the prosody requirement.

Performance evaluation is carried out after the implementation of the proposed strategies. It includes both objective test and subjective perceptual test. Results in the objective test showed that selected segments in proposed system are more suitable and better matched with the target prosody, in terms of pitch levels and segment durations. A better matched in acoustic properties means that it would require a lower degree of prosodic modification in later stage and the synthesized speech will be closer to the original human speech. Subjective perceptual test showed that our proposed system performs much better than the baseline system, with an absolute increase of about 0.4

to 0.5 marks in 5-point MOS. Some testing samples have absolute improvement of 1 mark or more. This means that the proposed system is much more intelligible and natural than the baseline, from the system users' point of view. However, there are still a small number of cases in which performance degradation occurred. We believe that the perceptual quality is not solely related to the acoustical parameters only.

7.2 Suggested future works

Several suggestions are made in order to further improve the Cantonese TTS system. Details are as follows:

1) Design a new acoustic inventory for corpus-based Cantonese TTS

   It was shown that the existing speech inventory contains only one unique copy for a large number of sub-syllable units. To facilitate a more useful unit selection process, we suggest developing a new speech inventory. In the new speech inventory, each required synthesis unit should contain several copies. We can choose different segments according to the conditions required in each speech utterance. Moreover, characters embedded in different words and phrases may contain different acoustic properties and co-articulations. In order to further improve the selection process of speech segments, such parameters are required in new speech inventory in order to have a better selection process.

2) Investigate the relationship between perceptual quality and acoustic parameters

So far we assume that the pitch level of the speech segment is the most important parameter to be considered in unit selection and concatenation. Other acoustic factors, including duration and intensity levels, may also affect the perceptual quality. We think the importance of different acoustic parameters may vary in different characters and in different speeches. So we can investigate the relationship between acoustic parameters and the perceptual quality of the synthesized speech.

References:

[1] K.M. Law, "Cantonese Text-to-Speech Synthesis Using Sub-syllable Units", M. Phil. Thesis, the Chinese University of Hong Kong, 2001

Appendix 1

Mean pitch level of Initials and Finals stored in the inventory

Long Vowels

Final units	Tone	Count	Mean	SD	Max	Min
aa	ALL	376	182.2	23.9	266.7	131.0
	1	45	232.4	15.6	266.7	202.5
	2	24	160.9	12.1	183.9	142.9
	3	184	184.4	8.5	205.1	164.9
	4	17	153.9	14.4	186.0	131.0
	5	64	160.9	11.5	225.4	144.1
	6	42	174.8	12.3	202.5	150.9
e	ALL	149	180.5	21.6	238.8	136.8
	1	10	227.7	7.7	238.8	216.2
	2	17	163.3	11.9	190.5	141.6
	3	56	194.0	12.1	216.2	168.4
	4	4	156.4	12.2	172.1	142.9
	5	55	166.9	10.2	181.8	136.8
	6	7	167.5	10.6	181.8	153.8
i	ALL	477	203.8	32.1	291.0	127.0
	1	215	234.3	12.3	291.0	207.8
	2	47	167.4	16.7	219.2	146.8
	3	57	197.5	12.9	219.2	161.6
	4	33	152.8	11.9	174.6	127.0
	5	61	181.0	14.5	219.2	139.1
	6	64	182.2	11.1	200.0	160.0
o	ALL	326	172.9	22.7	253.9	131.1
	1	19	228.5	15.2	253.9	205.1
	2	33	167.9	13.6	200.0	144.1
	3	72	193.8	11.9	219.2	172.0
	4	23	151.1	14.2	188.2	131.1
	5	168	161.9	11.1	190.5	142.9
	6	11	170.1	6.2	181.9	163.2

Final units	Tone	Count	Mean	SD	Max	Min
u	ALL	176	196.7	32.6	285.6	136.7
	1	43	244.2	15.2	285.6	219.2
	2	15	171.7	16.7	202.5	150.9
	3	52	197.0	14.7	222.2	152.4
	4	10	149.8	7.6	160.0	136.7
	5	48	175.4	10.6	202.5	155.4
	6	8	172.6	11.7	190.5	158.4
œ	ALL	102	182.7	16.3	231.9	153.8
	1	2	230.2	2.3	231.9	228.6
	2	49	168.9	9.1	197.6	153.8
	3	50	194.3	6.7	210.5	175.8
	4	1	188.3	0.0	188.3	188.3
	5	0	0.0	0.0	0.0	0.0
	6	0	0.0	0.0	0.0	0.0
yu	ALL	168	179.8	23.9	266.7	132.2
	1	8	235.0	17.0	266.7	210.5
	2	9	167.5	15.1	190.5	148.1
	3	52	198.9	9.0	213.4	166.7
	4	35	153.3	12.2	203.3	132.2
	5	59	173.7	13.4	195.1	141.6
	6	5	172.3	7.2	179.8	161.6

Diphthongs

Final units	Tone	Count	Mean	SD	Max	Min
aai	ALL	420	181.6	17.0	246.2	135.6
	1	62	206.7	10.2	246.2	183.9
	2	17	152.0	11.0	181.8	135.6
	3	191	184.8	9.4	213.4	160.0
	4	7	146.7	7.2	156.9	136.8
	5	54	163.6	6.4	179.8	145.5
	6	89	176.5	10.2	200.0	155.3
aau	ALL	300	176.0	20.2	254.0	130.1
	1	13	224.4	15.9	254.0	205.2
	2	13	163.7	18.3	200.0	144.2
	3	174	185.1	9.0	213.4	163.3
	4	44	146.0	7.3	163.3	130.1
	5	50	162.3	5.4	175.8	149.5
	6	6	165.6	12.2	188.3	152.4
iu	ALL	203	187.8	26.4	266.7	140.4
	1	34	234.6	13.8	266.7	205.1
	2	27	174.5	12.3	202.6	155.3
	3	64	192.0	12.6	235.2	161.6
	4	11	152.1	7.5	166.1	140.4
	5	58	171.1	10.2	192.8	142.9
	6	9	171.6	10.5	198.3	165.0
ui	ALL	133	181.1	18.6	262.3	138.0
	1	1	262.3	0.0	262.3	262.3
	2	12	158.8	10.2	181.8	146.8
	3	56	195.4	9.9	213.4	166.7
	4	4	145.1	8.2	156.9	138.0
	5	52	173.5	10.3	205.2	145.5
	6	8	171.0	10.4	188.3	155.3
eoi	ALL	191	178.5	22.4	271.2	140.4
	1	14	231.6	15.6	271.2	210.5
	2	34	163.4	14.0	205.1	144.1
	3	68	190.6	9.6	225.4	168.5
	4	7	150.1	7.9	163.3	140.4

Final units	Tone	Count	Mean	SD	Max	Min
oi	ALL	299	178.8	16.6	258.1	136.8
	1	10	226.4	16.1	258.1	207.8
	2	14	153.6	10.7	179.8	141.6
	3	177	183.9	9.5	219.2	165.0
	4	15	147.1	7.6	163.3	136.8
	5	47	171.9	8.3	192.8	152.4
	6	36	172.4	9.6	188.3	146.8
ai	ALL	445	191.0	35.5	275.8	131.1
	1	177	228.8	12.8	275.8	202.5
	2	20	157.0	6.9	173.9	146.8
	3	82	189.1	14.4	213.4	146.8
	4	93	146.6	9.1	188.3	131.1
	5	54	168.0	10.3	192.8	141.6
	6	19	165.5	6.4	181.8	152.4
au	ALL	464	187.9	32.2	291.0	132.2
	1	165	226.6	12.3	291.0	200.0
	2	58	167.0	13.8	210.6	140.4
	3	60	183.1	10.4	207.8	160.0
	4	54	149.8	8.3	166.7	132.2
	5	67	162.1	11.2	197.5	144.2
	6	60	170.2	13.0	219.2	145.5
ei	ALL	349	177.8	29.8	254.0	134.5
	1	64	228.0	11.4	254.0	200.0
	2	27	165.0	13.6	200.0	148.1
	3	68	191.5	11.7	225.4	169.0
	4	91	149.2	7.5	173.9	134.5
	5	75	163.0	10.2	186.0	137.9
	6	24	173.4	14.4	222.2	156.9
ou	ALL	505	176.3	24.4	275.8	130.1
	1	24	245.8	16.6	275.8	205.1
	2	39	164.6	16.8	197.5	135.6
	3	189	188.9	12.8	225.3	161.6
	4	96	150.5	9.4	202.6	130.1

	5	56	164.8	8.8	200.0	144.1
	6	12	171.5	10.2	188.2	153.8

	5	60	172.2	12.3	200.0	144.1
	6	97	167.4	10.9	205.1	146.8

Vowel with Nasal Codas

Final units	Tone	Count	Mean	SD	Max	Min
aam	ALL	306	203.8	31.0	266.7	135.6
	1	178	226.4	13.0	266.7	200.0
	2	4	154.6	8.1	160.0	142.9
	3	56	192.1	8.1	210.5	170.2
	4	17	148.4	7.8	164.9	135.6
	5	50	160.5	8.1	186.1	142.9
	6	1	156.9	0.0	156.9	156.9
aan	ALL	352	165.1	25.4	250.0	130.1
	1	29	225.0	13.7	250.0	200.0
	2	27	165.8	12.6	183.9	146.8
	3	58	189.9	8.4	205.1	170.2
	4	158	145.6	6.4	172.0	130.1
	5	58	162.3	8.4	190.5	141.6
	6	22	168.0	9.4	186.0	140.4
aang	ALL	266	174.9	16.5	250.1	137.9
	1	9	224.5	15.4	250.1	207.8
	2	2	159.7	20.1	173.9	145.5
	3	50	195.0	7.1	205.1	170.2
	4	1	139.1	0.0	139.1	139.1
	5	56	162.2	7.1	172.1	137.9
	6	148	170.4	7.6	192.8	150.9
im	ALL	127	185.4	19.3	250.0	135.6
	1	4	238.5	13.4	250.0	219.2
	2	7	173.0	16.0	197.6	152.4
	3	52	198.2	8.2	213.3	175.8
	4	4	141.7	4.4	145.5	135.6
	5	52	174.7	9.2	195.1	148.1
	6	8	177.3	17.0	205.1	153.9
in	ALL	252	189.1	35.1	285.7	132.2
	1	49	249.1	20.8	285.7	207.8

Final units	Tone	Count	Mean	SD	Max	Min
am	ALL	285	189.9	24.5	266.7	130.1
	1	40	236.5	17.2	266.7	190.5
	2	7	168.3	16.6	192.8	152.4
	3	169	189.4	10.8	222.2	166.7
	4	14	151.0	11.0	163.3	130.1
	5	51	170.2	8.4	192.8	149.5
	6	4	169.4	4.4	173.9	163.3
an	ALL	514	173.0	34.1	280.7	132.2
	1	92	234.5	15.2	280.7	202.6
	2	27	168.1	11.3	197.5	150.9
	3	66	195.3	12.4	222.2	170.2
	4	264	148.4	7.2	175.8	132.2
	5	57	164.0	7.7	181.8	144.1
	6	8	172.2	11.1	188.3	156.9
ang	ALL	290	203.2	29.6	266.7	131.1
	1	160	226.6	10.3	266.7	205.1
	2	53	165.3	7.7	177.8	148.1
	3	56	190.8	8.9	213.3	175.8
	4	17	150.7	12.8	183.9	131.1
	5	0	0.0	0.0	0.0	0.0
	6	4	165.2	7.1	175.8	161.6
ing	ALL	270	178.6	31.2	280.4	134.4
	1	39	239.2	16.9	280.4	210.5
	2	13	169.3	17.6	210.5	148.1
	3	65	189.5	10.1	219.2	168.4
	4	81	148.5	8.8	188.3	134.4
	5	55	172.3	9.6	190.5	136.8
	6	17	169.8	5.8	177.8	158.4
ung	ALL	378	195.6	32.9	307.7	139.1
	1	89	245.7	21.8	307.7	188.3

	2	14	167.3	7.7	183.9	158.4
	3	70	193.7	10.4	222.2	170.2
	4	46	151.7	10.1	179.0	132.2
	5	53	173.9	8.6	188.3	153.8
	6	20	168.0	7.2	183.9	155.4
on	ALL	245	184.0	17.7	254.0	144.2
	1	16	234.4	14.1	254.0	207.8
	2	1	150.9	0.0	150.9	150.9
	3	174	184.9	8.5	213.3	164.9
	4	2	149.0	6.9	153.8	144.2
	5	49	168.0	7.3	186.0	152.4
	6	3	162.7	11.5	170.2	149.5
un	ALL	139	184.6	22.8	258.1	140.3
	1	5	244.2	10.1	258.1	231.9
	2	10	163.6	8.4	183.9	156.9
	3	58	200.2	12.1	216.2	166.6
	4	16	153.0	9.3	175.8	140.3
	5	48	175.4	9.4	197.6	152.4
	6	2	165.0	4.8	168.4	161.6
eon	ALL	124	178.2	19.4	258.1	142.9
	1	3	244.5	17.1	258.1	225.4
	2	6	158.1	6.0	166.7	149.5
	3	53	192.6	7.9	207.8	173.9
	4	5	150.2	5.3	155.4	142.9
	5	47	165.6	7.9	188.3	152.4
	6	10	166.8	4.8	177.8	160.3

	2	17	169.0	17.6	205.1	149.5
	3	171	189.7	9.4	216.2	170.2
	4	32	152.3	8.1	170.2	139.1
	5	51	172.6	10.1	190.5	148.2
	6	18	171.2	7.8	186.0	158.4
eng	ALL	124	175.5	21.2	262.3	139.1
	1	4	235.6	20.0	262.3	216.2
	2	10	155.7	10.8	170.2	139.1
	3	54	190.5	8.3	207.8	170.2
	4	3	145.5	3.5	149.5	142.9
	5	47	160.0	9.6	200.0	142.9
	6	6	169.3	8.2	183.9	161.6
ong	ALL	332	184.7	24.8	285.7	139.0
	1	50	230.5	19.3	285.7	202.6
	2	20	158.2	11.8	195.2	148.2
	3	177	184.9	8.5	205.1	164.9
	4	27	151.0	7.4	168.4	139.0
	5	52	168.7	8.7	192.8	148.1
	6	6	173.3	11.4	190.5	160.0
oeng	ALL	186	178.4	26.9	262.3	132.2
	1	22	234.3	18.0	262.3	207.8
	2	6	166.9	25.8	216.2	149.5
	3	51	188.3	8.5	205.1	161.6
	4	21	145.5	9.3	177.8	132.2
	5	61	163.8	9.6	181.8	136.7
	6	25	175.0	9.5	190.5	158.0
yun	ALL	175	176.9	22.2	250.0	135.6
	1	8	233.4	10.6	250.0	219.2
	2	10	166.2	17.9	200.0	142.9
	3	55	194.4	8.4	207.8	168.4
	4	39	153.1	9.4	173.9	135.6
	5	55	172.1	10.9	195.1	142.8
	6	8	163.5	6.9	173.9	156.9

Vowel with Stop Codas

Appendix 1 Mean pitch level of Initials and Finals stored in the inventory

Final units	Tone	Count	Mean	SD	Max	Min
aap	ALL	180	187.7	11.1	235.3	158.4
	1	0	0.0	0.0	0.0	0.0
	2	0	0.0	0.0	0.0	0.0
	3	171	188.7	10.4	235.3	163.3
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	9	169.4	8.5	181.8	158.4
aat	ALL	239	184.4	11.9	242.4	146.8
	1	1	242.4	0.0	242.4	242.4
	2	48	173.4	10.0	205.1	146.8
	3	189	186.9	9.9	213.4	161.6
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	1	168.4	0.0	168.4	168.4
aak	ALL	247	173.3	17.1	246.1	146.8
	1	2	228.3	25.2	246.1	210.5
	2	162	164.5	11.2	231.9	146.8
	3	72	191.2	10.5	219.2	166.7
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	11	175.5	11.4	197.5	164.0
ek	ALL	122	183.6	16.3	258.0	145.5
	1	1	258.0	0.0	258.0	258.0
	2	48	169.5	9.4	192.8	145.5
	3	59	195.1	8.5	213.3	175.8
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	14	178.6	7.5	190.5	168.4
ip	ALL	116	197.8	14.0	228.6	158.2
	1	0	0.0	0.0	0.0	0.0
	2	48	194.2	12.8	222.2	158.2
	3	59	202.9	11.7	228.6	172.1
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	9	183.8	19.0	228.6	166.7

Final units	Tone	Count	Mean	SD	Max	Min
ut	ALL	60	198.4	14.7	225.4	153.8
	1	0	0.0	0.0	0.0	0.0
	2	0	0.0	0.0	0.0	0.0
	3	55	201.5	11.0	225.4	170.2
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	5	165.3	8.6	175.8	153.8
yut	ALL	123	196.1	17.2	228.6	155.3
	1	0	0.0	0.0	0.0	0.0
	2	48	188.9	12.6	216.2	161.6
	3	62	206.5	13.5	228.6	177.8
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	13	173.1	11.2	197.5	155.3
ap	ALL	296	186.0	17.9	253.9	148.1
	1	18	234.0	10.8	253.9	219.2
	2	47	183.1	10.4	205.1	148.1
	3	49	198.3	9.3	216.2	179.8
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	182	178.7	11.6	219.2	152.4
at	ALL	543	215.0	39.5	530.5	142.9
	1	271	249.6	24.2	530.5	195.1
	2	51	179.0	11.8	210.5	142.9
	3	48	197.2	9.3	219.2	170.2
	4	1	153.8	0.0	153.8	153.8
	5	0	0.0	0.0	0.0	0.0
	6	172	176.5	8.6	200.0	145.5
ak	ALL	270	213.6	23.9	271.1	158.4
	1	164	229.9	12.4	271.1	202.6
	2	49	186.7	12.4	222.2	158.4
	3	3	189.2	11.6	196.7	175.8
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	54	189.6	13.3	250.0	166.7

it	ALL	94	195.1	19.6	266.7	149.5
	1	1	266.7	0.0	266.7	266.7
	2	0	0.0	0.0	0.0	0.0
	3	64	203.6	12.4	231.9	172.0
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	29	173.8	10.8	200.0	149.5
ok	ALL	159	182.2	11.5	217.4	150.9
	1	0	0.0	0.0	0.0	0.0
	2	48	174.5	8.7	188.2	150.9
	3	95	187.5	9.9	217.4	156.9
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	16	174.5	11.2	192.8	155.3
ot	ALL	52	197.6	9.1	213.3	166.7
	1	0	0.0	0.0	0.0	0.0
	2	0	0.0	0.0	0.0	0.0
	3	52	197.6	9.1	213.3	166.7
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	0	0.0	0.0	0.0	0.0
oek	ALL	112	189.2	14.6	231.9	152.4
	1	0	0.0	0.0	0.0	0.0
	2	48	182.2	14.8	216.2	152.4
	3	60	195.9	11.1	231.9	163.2
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	4	173.0	2.4	175.8	170.2

eot	ALL	116	189.3	24.3	271.2	155.3
	1	16	244.1	17.6	271.2	216.2
	2	48	180.0	10.8	202.5	155.3
	3	0	0.0	0.0	0.0	0.0
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	52	181.1	6.6	190.8	159.6
ik	ALL	130	211.1	26.9	285.7	160.0
	1	49	238.7	14.5	285.7	207.8
	2	0	0.0	0.0	0.0	0.0
	3	50	205.8	8.4	228.6	183.9
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	31	176.1	9.8	199.6	160.0
uk	ALL	319	214.0	30.8	285.7	129.0
	1	186	236.1	15.7	285.7	197.6
	2	51	187.0	18.3	242.4	156.9
	3	0	0.0	0.0	0.0	0.0
	4	0	0.0	0.0	0.0	0.0
	5	0	0.0	0.0	0.0	0.0
	6	82	180.5	15.3	228.6	129.0

Nasal Finals

Final units	Tone	Count	Mean	SD	Max	Min
m	ALL	162	181.9	10.8	205.1	156.9
	1	0	0.0	0.0	0.0	0.0
	2	0	0.0	0.0	0.0	0.0
	3	0	0.0	0.0	0.0	0.0

Final units	Tone	Count	Mean	SD	Max	Min
ng	ALL	173	181.5	10.7	204.5	156.9
	1	0	0.0	0.0	0.0	0.0
	2	0	0.0	0.0	0.0	0.0
	3	0	0.0	0.0	0.0	0.0

Appendix 1 Mean pitch level of Initials and Finals stored in the inventory

	4	0	0.0	0.0	0.0	0.0
	5	1	190.5	0.0	190.5	190.5
	6	161	181.8	10.8	205.1	156.9

	4	1	170.2	0.0	170.2	170.2
	5	11	182.4	11.7	197.5	156.9
	6	161	181.5	10.7	204.5	163.3

Appendix 2

Mean duration of Initials and Finals stored in the inventory

Initials

Units	Mean	Standard Deviation	Maximum Duration	Minimum Duration
b	0.085	0.029	0.197	0.018
d	0.075	0.027	0.202	0.021
g	0.087	0.029	0.179	0.026
p	0.112	0.033	0.199	0.049
t	0.114	0.034	0.240	0.029
k	0.114	0.035	0.235	0.050
gw	0.117	0.042	0.245	0.037
kw	0.135	0.028	0.205	0.051
z	0.102	0.028	0.201	0.035
c	0.136	0.034	0.242	0.055
zh	0.101	0.029	0.191	0.043
ch	0.132	0.034	0.218	0.079
f	0.118	0.033	0.235	0.038
s	0.131	0.034	0.234	0.044
sh	0.131	0.034	0.210	0.062
h	0.099	0.027	0.207	0.020
m	0.069	0.022	0.192	0.021
n	0.081	0.019	0.139	0.042
ng	0.094	0.024	0.175	0.045
j	0.070	0.025	0.190	0.023
w	0.098	0.036	0.221	0.024
l	0.073	0.023	0.174	0.026

Finals

Units	Mean	Standard Deviation	Maximum Duration	Minimum Duration
aa	0.290	0.105	0.55	0.071
e	0.234	0.077	0.437	0.097
i	0.234	0.100	0.495	0.052
o	0.256	0.103	0.447	0.082
u	0.190	0.079	0.437	0.067

oe	0.200	0.058	0.387	0.102
yu	0.188	0.090	0.404	0.065
aai	0.329	0.119	0.550	0.090
aau	0.352	0.096	0.522	0.100
ai	0.316	0.094	0.510	0.069
au	0.266	0.097	0.513	0.084
ei	0.270	0.090	0.486	0.072
eoi	0.246	0.091	0.473	0.104
iu	0.218	0.098	0.450	0.085
oi	0.308	0.133	0.553	0.095
ou	0.273	0.110	0.496	0.070
ui	0.195	0.078	0.453	0.090
aam	0.323	0.100	0.560	0.113
aan	0.318	0.094	0.561	0.105
aang	0.367	0.085	0.527	0.121
im	0.211	0.069	0.422	0.117
in	0.227	0.099	0.474	0.077
on	0.318	0.129	0.574	0.117
un	0.191	0.090	0.430	0.080
am	0.295	0.100	0.480	0.113
an	0.262	0.084	0.473	0.099
ang	0.312	0.085	0.501	0.106
eon	0.230	0.079	0.488	0.102
ing	0.224	0.082	0.420	0.082
ung	0.258	0.117	0.519	0.093
eng	0.266	0.079	0.515	0.092
ong	0.311	0.135	0.604	0.094
oeng	0.229	0.092	0.480	0.120
yun	0.218	0.101	0.457	0.077
aap	0.232	0.070	0.413	0.067
aat	0.256	0.095	0.477	0.105
aak	0.256	0.070	0.470	0.084
ek	0.207	0.060	0.301	0.088
ip	0.139	0.031	0.248	0.078
it	0.142	0.040	0.220	0.067
ok	0.160	0.040	0.300	0.094

ot	0.118	0.03	0.197	0.077
oek	0.151	0.036	0.257	0.072
ut	0.113	0.040	0.193	0.054
yut	0.139	0.065	0.384	0.066
ap	0.172	0.051	0.299	0.080
at	0.142	0.051	0.284	0.055
ak	0.213	0.097	0.430	0.070
eot	0.192	0.067	0.320	0.047
ik	0.147	0.032	0.250	0.078
uk	0.140	0.041	0.271	0.053
m	0.325	0.128	0.577	0.068
ng	0.324	0.124	0.510	0.075

Appendix 3

Mean intensity level of Initials and Finals stored in the inventory

Initials

Units	Mean	Standard Deviation	Maximum Duration	Minimum Duration
b	206.9	183.9	1306.6	6.6
d	217.0	187.2	1293.7	9.0
g	189.3	163.6	1910.0	27.4
p	398.9	349.2	4008.7	37.7
t	346.2	191.6	1036.6	62.1
k	264.7	156.1	934.3	52.9
gw	796.2	411.6	2706.0	80.3
kw	404.2	270.9	1643.0	68.4
z	177.2	121.7	1184.7	23.3
c	447.8	205.3	1284.5	117.6
zh	296.5	156.5	802.3	71.8
ch	639.4	206.8	1174.1	235.2
f	161.5	82.9	663.3	45.3
s	487.9	252.0	2409.9	125.1
sh	775.1	268.0	1656.1	371.2
h	383.8	281.5	1991.0	64.8
m	667.9	342.9	2239.5	138.1
n	993.5	463.7	2186.1	159.9
ng	1121.1	540.6	2803.3	129.0
j	683.7	476.4	3353.8	40.1
w	1104.7	661.9	5111.2	117.3
l	875.9	468.1	2966.5	144.6

Finals

Units	Mean	Standard Deviation	Maximum Duration	Minimum Duration
aa	1374.1	496.7	3328.3	386.6
e	1554.9	481.1	2777.6	379.1
i	1013.8	382.9	2924.2	179.1

o	1295.8	563.3	3529.5	235.9
u	1482.3	532.4	2812.9	173.6
oe	1627.8	551.3	5560.4	885.0
yu	1053.9	443.0	2294.1	181.3
aai	1487.4	463.3	3499.2	248.2
aau	1278.2	489.5	4217.8	197.7
ai	1279.2	539.9	3899.5	246.6
au	1320.2	488.4	3563.8	309.6
ei	1248.6	641.4	5085.2	257.6
eoi	1434.8	604.8	5021.5	360.4
iu	1528.5	519.4	3033.4	360.4
oi	1323.8	470.0	4265.9	237.6
ou	1290.1	576.5	4125.5	207.1
ui	1654.5	513.6	2884.5	392.0
aam	1218.8	373.4	2721.6	361.9
aan	924.2	454.5	3019.5	219.5
aang	897.4	259.5	1904.7	239.0
im	1001.0	373.3	2238.0	378.5
in	1069.3	395.1	1950.7	275.5
on	1236.3	353.6	2891.8	481.6
un	1258.1	438.3	2271.0	248.7
am	956.4	304.5	2251.6	371.8
an	900.4	432.6	2749.9	157.8
ang	881.4	322.2	2095.6	348.0
eon	1105.6	388.7	3163.4	427.1
ing	1034.2	473.6	2055.0	205.8
ung	1129.3	431.1	3153.7	274.3
eng	1170.4	338.5	2285.7	401.3
ong	1344.9	446.1	2615.9	385.6
oeng	1165.4	399.7	2236.4	305.7
yun	1315.0	587.3	2328.0	307.0
aap	1217.2	376.1	2449.1	424.1
aat	1175.2	478.3	2751.9	323.2
aak	1032.2	372.8	2993.9	284.7
ek	830.1	280.0	1543.4	231.5
ip	770.3	284.0	2216.6	307.3

Appendix 3 Mean intensity level of Initials and Finals stored in the inventory

it	827.5	354.6	1959.5	247.3
ok	1241.2	418.7	4000.5	501.0
ot	1235.9	414.1	1995.7	471.8
oek	1091.2	335.5	2685.3	487.9
ut	1055.4	294.0	1728.3	509.4
yut	1041.9	312.5	1818.3	419.3
ap	841.1	475.0	2545.5	98.2
at	1156.4	514.8	4433.3	172.7
ak	908.4	522.1	3340.0	184.4
eot	824.0	426.8	2564.5	285.2
ik	797.8	502.7	3465.4	148.2
uk	898.0	394.5	2639.9	149.7
m	782.2	643.7	7768.2	214.0
ng	752.3	263.0	1889.7	256.5

Appendix 4

Test word used in performance evaluation

1	交換	gaau1-wun6
2	德州	dak1-zau1
3	酒樓	zau2-lau4
4	彎曲	waan1-kuk1
5	湖泊	wu4-pok3
6	華潤	waa4-jeon6
7	沙田	saa1-tin4
8	待遇	doi6-jyu6
9	上環	soeng6-waan4
10	花朵	faa1-do2
11	報名	bou3-meng2
12	禮貌	lai5-maau6
13	字典	zi6-din2
14	力量	lik6-loeng6
15	意願	ji3-jyun6
16	合約	hap6-joek3
17	渴望	hot3-mong6
18	作洞	zok3-dung6
19	翻案	faan1-ngon3
20	大浪	daai6-long6
21	落車	lok6-ce1
22	交通	gaau1-tung1
23	銀行	ngan4-hong4
24	信件	seon3-gin2
25	里昂	lei5-ngong4

26	不屈	bat1-wat1
27	惠康	wai6-hong1
28	區域	keoi1-wik6
29	風雨	fung1-jyu5
30	藍田	laam4-tin4
31	巴士	baa1-si2
32	彩虹	coi2-hung4
33	客戶	haak3-wu6
34	內容	noi6-jung4
35	光波	gwong1-bo1
36	收音機	sau1-jam1-gei1
37	奧克蘭	ou3-hak1-laan4
38	維他命	wai4-taa1-ming6
39	壞朋友	waai6-pang4-jau5
40	新世界	san1-sai3-gaai3
41	油麻地	jau4-maa4-dei6
42	實驗室	sat6-jim6-sat1
43	光碟機	gwong1-dip2-gei1
44	當事人	dong1-si6-jan4
45	飛機場	fei1-gei1-coeng4
46	香格里拉	hoeng1-gaak3-lei5-laai1
47	花鳥蟲魚	faa1-niu5-cung4-jyu2
48	郊野公園	gaau1-je5-gung1-jyun2
49	世界地圖	sai3-gaai3-dei6-tou4
50	文房四寶	man4-fong4-sei3-bou2

Appendix 5

Test paragraph used in performance evaluation

1	香港迪士尼每日可接待約三萬名遊客。
	hoeng1-gong2-dik6-si6-nei4-mui5-jat6-ho2-zip3-doi6-joek3-saam1-maan6-ming4-jau4-haak3
2	埃克森美孚調高先力超低硫柴油售價。
	ngaai1-hak1-sam1-mei5-ful-tiu4-goul-sin1-lik6-ciul-dai1-lau4-caai4-jau4-sau6-gaa3
3	北京用電負荷突破歷史紀錄。
	bak1-ging1-jung6-din6-fu6-ho6-dat6-po3-lik6-si2-gei2-luk6
4	十五個食物樣本不合格。
	sap6-ng5-go3-sik6-mat6-joeng6-bun2-bat1-hap6-gaak3
5	京九鐵路龍川段列車延誤。
	ging1-gau2-tit3-lou6-lung4-cyun1-dyun6-lit6-ce1-jin4-ng6
6	令人憂慮新聞自由會受影響。
	ling6-jan4-jau1-leoi6-san1-man4-zi6-jau4-wui6-sau6-jing2-hoeng2
7	各部門同意由地政處查核誰應負責維修護土牆。
	gok3-bou6-mun4-tung4-ji3-jau4-dei6-zing3-cyu5-caa4-hat6-seoi4-jing3-fu6-zaak3-wai4-sau1-wu6-tou2-coeng4
8	她就事件保持沈默而受到外界的指摘。
	taa1-zau6-si6-gin2-bou2-ci4-cam4-mak6-ji4-sau6-dou3-ngoi6-gaai3-dik1-zi2-zaak6
9	申請人因病情惡化去世。
	san1-cing2-jan4-jan1-beng6-cing4-ngok3-faa3-heoi3-sai3
10	南韓國防部長尹光雄向總統盧武鉉請辭。
	naam4-hon4-gwok3-fong4-bou6-zoeng2-wan5-gwong1-hung4-hoeng3-zung2-tung2-lou4-mou5-jyun5-cing2-ci4
11	女藝人應采兒被控藏毒案在九龍城裁判法院聆訊。
	neoi5-ngai6-jan4-jing3-coi2-ji4-bei6-hung3-cong4-duk6-ngon3-zoi6-gau2-lung4-sing4-coi4-pun3-faat3-jyun2-ling4-seon3
12	警方新界北交通意外特別調查隊跟進調查意外原因。
	ging2-fong1-san1-gaai3-bak1-gaau1-tung1-ji3-ngoi6-dak6-bit6-tiu4-caa4-deoi6-gan1-zeon3-tiu4-caa4-ji3-ngoi6-jyun4-jan1

13	活躍的西南季候風為華南帶來不穩定的天氣。
	wut6-joek3-dik1-sai1-naam4-gwai3-hau6-fung1-wai6-waa4-naam4-daai3-loi4-bat1-wan2-ding6-dik1-tin1-hei3
14	巴人自治政府主席阿巴斯的車隊離開沙龍在耶路撒冷的總理府。
	baa1-jan4-zi6-zi6-zing3-fu2-zyu2-zik6-aa3-baa1-si1-dik1-ce1-deoi2-lei4-hoi1-saa1-lung4-zoi6-je4-lou6-saat3-laang5-dik1-zung2-lei5-fu2
15	兩名被告分別將各自擁有的一個愉景灣複式洋房單位及一個九龍灣工業大廈單位出租。
	loeng5-ming4-bei6-gou3-fan1-bit6-zoeng1-gok3-zi6-jung2-jau5-dik1-jat1-go3-jyu4-ging2-waan1-fuk1-sik1-joeng4-fong4-daan1-wai2-kap6-jat1-go3-gau2-lung4-waan1-gung1-jip6-daai6-haa6-daan1-wai2-ceot1-zou1
16	災害發生後，廣梅汕鐵路總公司及時啟動了防洪搶險預案。
	zoi1-hoi6-faat3-sang1-hau6 gwong2-mui4-saan3-tit3-lou6-zung2-gung1-si1-kap6-si4-kai2-dung6-liu5-fong4-hung4-coeng2-him2-jyu6-ngon3
17	該官員還指出，中海油仍維持高增長及財務穩健等策略。
	goi1-gun1-jyun4-waan4-zi2-ceot1 zung1-hoi2-jau4-jing4-wai4-ci4-gou1-zang1-zoeng2-kap6-coi4-mou6-wan2-gin6-dang2-caak3-loek6
18	其中住宅用途的建築樓面面積有二十萬三千四百零四平方米，共三千一百三十一個單位。
	kei4-zung1-zyu6-zaak6-jung6-tou4-dik1-gin3-zuk1-lau4-min2-min6-zik1-jau5-ji6-sap6-maan6-saam1-cin1-sei3-baak3-ling4-sei3-ping4-fong1-mai5 gung6-saam1-cin1-jat1-baak3-saam1-sap6-jat1-go3-daan1-wai2
19	旺角長旺道一名外籍女傭被高空墜下鋁窗擊中頭部，受重傷送院救治。
	wong6-gok3-coeng4-wong6-dou6-jat1-ming4-ngoi6-zik6-neoi5-jung4-bei6-gou1-hung1-zeoi6-haa6-leoi5-coeng1-gik1-zung3-tau4-bou6 sau6-cung5-soeng1-sung3-jyun2-gau3-zi6
20	民間人權陣線今年七月一日繼續發起遊行，今年的主題是爭取全面普選及反對官商勾結。
	man4-gaan1-jan4-kyun4-zan6-sin3-gam1-nin4-cat1-jyut6-jat1-jat6-gai3-zuk6-faat3-hei2-jau4-hang4 gam1-nin4-dik1-zyu2-tai4-si6-zang1-ceoi2-cyun4-min6-pou2-syun2-kap6-faan2-deoi3-gun1-soeng1-ngau1-git3

21	美國總統布殊宣布接受越南總理潘文凱邀請，期待明年訪問越南，並參加在越南舉辦的亞太經合組織首腦會議。
	mei5-gwok3-zung2-tung2-bou3-syu4-syun1-bou3-zip3-sau6-jyut6-naam4-zung2-lei5-pun1-man4-hoi2-jiu1-cing2 kei4-doi6-ming4-nin4-fong2-man6-jyut6-naam4 bing6-caam1-gaa1-zoi6-jyut6-naam4-geoi2-baan6-dik1-ngaa3-taai3-ging1-hap6-zou2-zik1-sau2-nou5-wui6-ji5
22	清晨五時許，保安發現寵物店的玻璃門粉碎，於是立即報警。
	cing1-san4-ng5-si4-heoi2 bou2-on1-faat3-jin6-cung2-mat6-dim3-dik1-bo1-lei1-mun4-fan2-seoi3 jyu1-si6-lap6-zik1-bou3-ging2
23	屋宇署於今年五月共批出二十份建築圖則，其中港島十一份、九龍三份及新界六份。
	nguk1-jyu5-cyu3-jyu1-gam1-nin4-ng5-jyut6-gung6-pai1-ceot1-ji6-sap6-fan6-gin3-zuk1-tou4-zak1 kei4-zung1-gong2-dou2-sap6-jat1-fan6 gau2-lung4-saam1-fan6-kap6-san1-gaai3-luk6-fan6
24	昨日下午，四川省相關部門專家組成的爆炸事故調查組，對事故展開全面調查。
	zok3-jat6-haa6-ng5 sei3-cyun1-saang2-soeng1-gwaan1-bou6-mun4-zyun1-gaa1-zou2-sing4-dik1-baau3-zaa3-si6-gu3-tiu4-caa4-zou2 deoi3-si6-gu3-zin2-hoi1-cyun4-min6-tiu4-caa4
25	中國常駐聯合國代表在聯大表明，中國反對為安理會改革人為設定時限，也反對強行表決有重大分歧的方案。
	zung1-gwok3-soeng4-zyu3-lyun4-hap6-gwok3-doi6-biu2-zoi6-lyun4-daai6-biu2-ming4 zung1-gwok3-faan2-deoi3-wai6-on1-lei5-wui2-goi2-gaak3-jan4-wai4-cit3-ding6-si4-haan6 jaa5-faan2-deoi3-koeng5-hang4-biu2-kyut3-jau5-zung6-daai6-fan1-kei4-dik1-fong1-ngon3

Appendix 6

Pitch profile used in the Text-to-Speech system

The following table is the pitch profile for different tone classes in Cantonese. There are 24 values for each tone class to show the ideal pitch changing among the syllable unit. When calculating the difference between it and the selected units, its mean values will be compared.

Values	Pitch levels (Hz)					
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6
1	246.3	181.0	208.9	191.8	191.0	199.6
2	245.6	179.1	207.2	189.0	188.7	197.8
3	244.9	177.5	205.7	186.2	186.4	196.0
4	244.5	176.1	204.3	183.4	184.3	194.4
5	244.1	175.1	203.1	180.8	182.4	192.9
6	243.8	174.3	202.0	178.1	180.8	191.5
7	243.6	173.9	201.0	175.6	179.4	190.1
8	243.4	173.7	200.2	173.2	178.3	188.9
9	243.4	173.8	199.4	170.9	177.6	187.9
10	243.4	174.2	198.8	168.8	177.0	187.0
11	243.4	175.1	198.3	166.8	176.8	186.2
12	243.5	176.3	197.8	165.1	177.0	185.6
13	243.6	177.9	197.6	163.5	177.3	185.1
14	243.8	180.2	197.4	162.2	178.2	184.7
15	244.0	183.0	197.5	161.2	179.3	184.4
16	244.2	186.3	197.6	160.5	180.8	184.3
17	244.3	190.3	197.7	160.1	182.6	184.4
18	244.2	194.6	197.8	160.0	184.8	184.5
19	243.8	199.1	197.8	160.2	187.1	184.6
20	242.8	203.5	197.7	160.5	189.4	184.6
21	240.9	207.1	197.4	160.8	191.1	184.5
22	238.0	209.4	196.8	161.4	192.3	184.5
23	235.0	210.7	196.6	162.5	193.2	184.7
24	232.8	211.8	196.9	163.7	194.2	185.2
MEAN VALUE	242.8	186.0	199.8	169.4	183.8	188.1

Appendix 7

Duration model used in Text-to-Speech system

Syllable	Initial duration (ms)	Final duration (ms)
aa	0	170
aai	0	180
aak	0	130
aam	0	190
aan	0	230
aang	0	180
aap	0	150
aat	0	140
aau	0	160
ai	0	190
ak	0	70
am	0	230
an	0	170
ang	0	160
ap	0	60
at	0	70
au	0	180
baa	32	160
baai	32	170
baak	32	100
baan	32	190
baang	32	170
baat	40	100
baau	32	170
bai	32	160
bak	32	70
bam	24	190
ban	24	190
bang	24	170
bat	32	70
bau	32	160

Syllable	Initial duration (ms)	Final duration (ms)
be	32	170
bei	24	150
bek	32	70
beng	24	180
bi	32	120
bik	32	70
bin	24	160
bing	32	160
bit	32	70
biu	32	150
bo	32	160
bok	32	110
bong	32	200
bou	24	160
bui	32	150
buk	32	60
bun	32	170
bung	32	160
but	40	60
caa	70	140
caai	80	160
caak	70	90
caam	70	180
caan	70	180
caang	70	180
caap	70	70
caat	70	100
caau	85	160
cai	70	120
cak	70	140
cam	70	130

Syllable	Initial duration (ms)	Final duration (ms)
can	70	160
cang	70	150
cap	70	60
cat	70	60
cau	70	130
ce	60	150
cek	60	90
ceng	60	170
ceoi	60	120
ceon	70	140
ceot	70	60
ci	70	120
cik	70	70
cim	70	140
cin	70	140
cing	70	150
cip	70	60
cit	70	60
ciu	70	120
co	70	160
coe	70	140
coek	70	80
coeng	70	160
coi	70	140
cok	70	80
cong	70	170
cou	70	150
cuk	70	70
cung	70	160
cyu	70	130
cyun	80	140

Appendix 7 Duration model used in Text-to-Speech system

Syllable	Initial duration (ms)	Final duration (ms)
cyut	80	60
daa	32	160
daai	32	160
daam	32	197
daan	32	190
daap	40	90
daat	40	100
dai	32	150
dak	32	70
dam	32	170
dan	40	170
dang	32	170
dap	32	60
dat	32	60
dau	32	160
de	32	170
dei	24	160
dek	24	70
deng	32	200
deoi	32	150
deon	32	200
di	32	120
dik	24	60
dim	24	170
din	32	180
ding	32	160
dip	24	90
dit	24	80
diu	32	140
do	32	150
doe	32	160
doek	40	100
doeng	32	160
doi	32	160
dok	32	110
don	32	190

Syllable	Initial duration (ms)	Final duration (ms)
dong	32	190
dou	32	150
duk	32	60
dung	32	160
dyun	40	170
dyut	32	80
faa	80	160
faai	72	160
faak	80	90
faan	80	190
faat	72	100
fai	80	150
fan	64	160
fat	72	70
fau	64	160
fe	64	170
fei	80	140
fi	80	110
fing	80	140
fo	80	150
fok	64	100
fong	64	180
fu	80	140
fui	96	120
fuk	72	60
fun	80	170
fung	72	170
fut	88	80
gaa	32	160
gaai	40	170
gaak	40	100
gaam	40	200
gaan	40	200
gaang	40	170
gaap	40	90
gaat	40	100

Syllable	Initial duration (ms)	Final duration (ms)
gaau	40	180
gai	40	140
gak	40	70
gam	32	170
gan	40	170
gang	40	180
gap	40	60
gat	40	60
gau	40	150
ge	40	160
gei	40	160
geng	40	220
geoi	32	160
gik	40	60
gim	32	160
gin	32	170
ging	40	170
gip	48	70
git	48	70
giu	48	150
go	32	150
goe	32	160
goek	40	90
goeng	40	160
goi	40	160
gok	40	100
gon	40	190
gong	40	200
got	40	110
gou	40	160
gu	40	150
gui	40	140
guk	40	90
gun	40	170
gung	40	160
gwaa	40	170

Appendix 7 Duration model used in Text-to-Speech system

Syllable	Initial duration (ms)	Final duration (ms)
gwaai	32	160
gwaak	40	100
gwaan	32	200
gwaat	40	100
gwai	40	160
gwan	40	180
gwang	48	180
gwat	48	70
gwik	48	80
gwing	48	170
gwo	40	150
gwok	32	110
gwong	48	200
gwun	48	170
gyun	40	180
gyut	48	80
haa	32	180
haai	32	200
haak	48	140
haam	32	230
haan	32	240
haang	48	200
haap	80	80
haat	72	80
haau	48	210
hai	40	180
hak	56	100
ham	64	180
han	48	210
hang	32	200
hap	40	80
hat	40	90
hau	40	210
he	40	180
hei	32	190
hek	40	150

Syllable	Initial duration (ms)	Final duration (ms)
heng	32	230
heoi	24	190
him	56	210
hin	32	230
hing	40	210
hip	48	110
hit	48	80
hiu	48	150
ho	48	190
hoe	80	190
hoeng	40	150
hoi	40	200
hok	48	110
hon	40	240
hong	50	180
hot	72	100
hou	32	190
huk	30	130
hung	32	210
hyun	32	210
hyut	48	120
i	0	190
jaa	40	190
jaai	40	200
jaam	32	240
jaang	32	240
jaap	32	90
jaau	32	240
jai	32	200
jam	40	220
jan	40	210
jap	32	110
jat	40	120
jau	40	200
je	40	200
jeng	32	220

Syllable	Initial duration (ms)	Final duration (ms)
jeoi	32	230
jeon	32	220
ji	40	190
jik	40	110
jim	40	220
jin	40	220
jing	40	210
jip	40	130
jit	40	150
jiu	40	200
jo	40	200
joe	32	180
joek	40	160
joeng	32	160
juk	40	110
jung	32	210
jyu	40	190
jyun	40	220
jyut	40	140
kaa	72	140
kaai	64	150
kaan	64	190
kaat	64	100
kaau	64	160
kai	96	150
kak	64	70
kam	64	150
kan	72	130
kang	64	150
kap	64	60
kat	56	60
kau	56	150
ke	56	150
kei	56	140
kek	56	80
keng	56	180

Appendix 7 Duration model used in Text-to-Speech system

Syllable	Initial duration (ms)	Final duration (ms)
keoi	48	140
kik	56	60
kim	56	140
kin	56	150
king	56	170
kit	56	70
kiu	56	130
koe	56	140
koek	56	100
koeng	56	160
koi	56	160
kok	56	90
kong	64	180
ku	56	130
kui	56	140
kuk	56	60
kung	56	160
kut	56	80
kwaa	80	160
kwaai	72	160
kwaak	72	90
kwaan	72	190
kwaang	80	180
kwai	80	140
kwan	80	160
kwang	72	170
kwik	48	80
kwok	56	110
kwong	80	180
kyun	72	130
kyut	72	70
laa	32	200
laai	32	210
laak	32	170
laam	32	240
laan	40	230

Syllable	Initial duration (ms)	Final duration (ms)
laang	40	250
laap	40	150
laat	32	170
laau	32	250
lai	32	200
lak	32	140
lam	40	210
lan	32	210
lang	32	220
lap	32	140
lat	32	140
lau	40	200
le	32	220
lei	40	190
lek	32	140
leng	32	270
leoi	32	200
leon	32	220
leot	40	140
li	32	190
lik	32	110
lim	32	230
lin	40	210
ling	32	210
lip	32	130
lit	40	130
liu	32	190
lo	40	190
loe	32	220
loek	40	170
loeng	40	150
loi	32	210
lok	32	150
lon	32	230
long	40	250
lou	40	200

Syllable	Initial duration (ms)	Final duration (ms)
luk	40	110
lung	32	220
lyun	40	210
lyut	32	140
m	32	220
maa	40	230
maai	40	220
maak	32	170
maan	40	250
maang	40	250
maat	32	170
maau	40	250
mai	32	230
mak	40	140
man	32	230
mang	32	200
mat	40	140
mau	40	230
me	32	220
mei	32	220
meng	32	250
mi	32	190
mik	32	110
min	32	230
ming	32	210
mit	32	150
miu	32	220
mo	32	200
mok	40	180
mon	32	250
mong	32	260
mou	40	220
mui	32	220
muk	40	130
mun	32	220
mung	32	240

Appendix 7 Duration model used in Text-to-Speech system

Syllable	Initial duration (ms)	Final duration (ms)
mut	32	140
naa	40	210
naai	32	210
naak	32	170
naam	32	240
naan	32	230
naap	40	140
naau	32	260
nai	32	190
nam	32	200
nan	32	210
nang	32	220
nap	32	140
nau	32	200
ne	32	220
nei	40	180
neng	32	250
neoi	40	200
neon	32	220
ng	48	190
ngaa	48	200
ngaai	32	200
ngaak	40	180
ngaam	32	240
ngaan	48	230
ngaang	32	260
ngaap	32	160
ngaat	48	160
ngaau	32	260
ngai	48	170
ngak	32	140
ngam	32	200
ngan	40	190
ngang	32	220
ngap	32	140
ngat	32	140

Syllable	Initial duration (ms)	Final duration (ms)
ngau	40	200
ngo	40	190
ngoi	48	220
ngok	32	150
ngon	48	230
ngong	32	240
ngou	48	190
nguk	40	110
ngung	32	200
nik	32	110
nim	32	230
nin	32	210
ning	32	200
nip	32	130
nit	32	130
niu	32	200
no	32	200
noe	32	220
noek	32	180
noeng	32	250
noi	32	230
nok	32	160
nong	32	250
nou	40	220
nuk	32	110
nung	40	210
nyun	32	210
nyut	32	140
o	0	190
oi	0	200
ok	0	140
on	0	220
ong	0	220
ou	0	170
paa	56	140
paai	56	141

Syllable	Initial duration (ms)	Final duration (ms)
paak	64	100
paan	64	180
paang	56	160
paau	64	160
pai	64	140
pan	64	160
pang	48	150
pat	56	60
pau	64	140
pe	64	150
pei	56	130
pek	56	80
peng	56	180
pik	64	60
pin	56	160
ping	48	150
pit	64	70
piu	56	140
po	56	150
pok	48	90
pong	64	160
pou	56	150
pui	72	130
puk	56	50
pun	70	170
pung	48	170
put	56	80
saa	88	150
saai	88	160
saak	88	90
saam	96	190
saan	88	190
saang	88	180
saap	88	70
saat	88	90
saau	80	180

Syllable	Initial duration (ms)	Final duration (ms)
sai	88	140
sak	80	50
sam	80	160
san	80	160
sang	100	170
sap	90	60
sat	72	60
sau	80	150
se	88	130
sei	96	110
sek	96	90
seng	80	180
seoi	88	170
seon	80	170
seot	80	70
si	80	150
sik	80	60
sim	104	160
sin	96	150
sing	80	170
sip	112	60
sit	96	60
siu	96	130
so	80	160
soe	88	130
soek	88	90
soeng	80	160
soi	88	140
sok	96	90
son	88	190
song	88	200
sou	96	140
suk	88	50
sung	88	160
syu	96	120
syun	80	150

Syllable	Initial duration (ms)	Final duration (ms)
syut	88	70
taa	48	160
taai	64	160
taam	64	170
taan	48	200
taap	72	100
taat	64	100
tai	48	130
tam	64	150
tan	40	170
tang	64	130
tap	64	60
tat	64	60
tau	56	150
te	64	150
tek	56	70
teng	48	190
teoi	64	130
teon	56	170
ti	64	110
tik	64	60
tim	72	140
tin	56	150
ting	48	150
tip	56	70
tit	56	70
tiu	56	110
to	56	150
toe	56	140
toi	56	140
tok	48	100
tong	56	160
tou	48	150
tuk	64	50
tung	48	160
tyun	48	150

Syllable	Initial duration (ms)	Final duration (ms)
tyut	56	80
uk	0	110
ung	0	160
waa	32	210
waai	32	210
waak	40	160
waan	32	240
waang	40	230
waat	40	170
wai	40	190
wan	40	230
wang	32	210
wat	40	90
we	40	200
wet	32	90
wi	40	190
wik	32	150
wing	40	210
wo	40	180
wok	40	160
wong	32	240
wu	40	190
wui	40	220
wun	32	250
wut	40	130
zaa	40	160
zaai	40	170
zaak	32	110
zaam	40	190
zaan	40	180
zaang	40	190
zaap	40	90
zaat	40	100
zaau	40	170
zai	40	140
zak	40	70

Appendix 7 Duration model used in Text-to-Speech system

Syllable	Initial duration (ms)	Final duration (ms)
zam	40	140
zan	40	170
zang	40	170
zap	40	60
zat	40	60
zau	40	150
ze	40	160
zek	48	90
zeng	48	190
zeoi	48	130
zeon	48	150
zeot	48	70

Syllable	Initial duration (ms)	Final duration (ms)
zi	40	130
zik	40	60
zim	40	170
zin	40	150
zing	40	160
zip	48	60
zit	48	70
ziu	48	130
zo	48	140
zoe	40	150
zoek	48	100
zoeng	40	180

Syllable	Initial duration (ms)	Final duration (ms)
zoi	40	150
zok	40	90
zon	40	170
zong	40	190
zou	48	150
zuk	40	60
zung	40	160
zyu	48	140
zyun	56	150
zyut	56	70